le cnam

ÉCOLE DOCTORALE Arts & Métiers

Laboratoire Génomique Bioinformatique et Applications (GBA)

THÈSE présentée par :

Lieng Taing

soutenue le : **27 septembre 2012**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline/ Spécialité : BioInformatique

Approches bioinformatiques pour l'exploitation des données génomiques

THÈSE dirigée par :

M. ZAGURY Jean François	Professeur, Conservatoire National des Arts et Métiers
M. DELANEAU Olivier	Docteur, Université d'Oxford

RAPPORTEURS :

Mme GUINOT Christiane	Docteur, CERIES
Mme MATHIEU Flavie	Docteur, Unité INSERM U958

JURY :

M. GUEDJ Mickaël	Docteur, Pharnext
M. LATOUCHE Aurélien	Professeur, Conservatoire National des Arts et Métiers

"J'imagine qu'il peut être vrai que
la fortune dispose de la moitié de nos actions
mais qu'elle en laisse à peu près l'autre moitié en notre pouvoir"
Machiavel - Le prince - Chap. XXV

Remerciements

J'exprime toute ma reconnaissance au Pr. Jean-François Zagury, mon directeur de thèse pour m'avoir donné ma chance, guidé au cours de ces années, soutenu dans mes travaux. Ces années ont été les plus riches en enseignement.

Je remercie le Dr. Olivier Delaneau d'avoir co-dirigé ma thèse, et notamment pour le temps qu'il a consacré à mes problématiques et ses conseils avisés en programmation.

Je suis très reconnaissant envers le Dr. Christiane Guinot d'avoir accepté de juger mon travail de thèse en tant que rapporteur.

J'exprime aussi toute ma reconnaissance au Dr. Flavie Mathieu pour avoir accepté de juger mon travail de thèse en tant que rapporteur malgré les délais contraignants.

Je remercie le Pr. Aurélien Latouche de m'avoir fait l'honneur de participer au jury de cette thèse.

Je remercie également le Dr. Mickaël Guedj de m'avoir fait l'honneur de participer au jury de cette thèse.

J'exprime toute ma gratitude pour le CNAM ainsi que son personnel pour m'avoir accueilli. Je remercie le ministère de l'éducation et de la recherche pour le financement qu'elle m'a accordé et espère m'en être montré digne. Je remercie aussi tous les patients, médecins et instituts qui m'ont fournis les données avec lesquelles j'ai pu travailler. Je remercie aussi la République Française sans qui je ne serais certainement pas là aujourd'hui.

Je tiens aussi à remercier le laboratoire Génomique Bioinformatique et Applications du CNAM pour son accueil, ses débats enflammés et autres péripéties.

Par ordre d'apparition : Cedcoul connu sous le sobriquet de M. Non, pour son soutien de tous les jours ; Kos pour ses réflexions poussées jusqu'à l'autisme, son assistance technique hors pair ; TFK pour sa sérénité, ses tweets ; Fofy pour sa passion envers les dessins animés de notre enfance, son professionnalisme frisant la névrose ; Sigrid pour son envie de "Vivre", son esprit contestataire ; Jean-Louis pour sa classe et sa Corse ; Roudz pour sa paranoïa, ses délicieux petits mets et Christiane pour sa disponibilité ainsi que nos discussions cinéphiles. Petit noyau ancestral sans qui je n'aurais probablement pas effectué ce long périple.

Viennent ensuite : Maitre Zu pour sa geekerie suprême et son hospitalité légendaire, LN pour sa geekerie modérée ainsi que son esprit acerbe, Hady pour ses goûts musicaux et sa bonne humeur, Mister Do! pour sa vision de la vie, Lu pour sa joie de vivre et sa franchise, Nesrine pour son humour et ses antihistaminiques, Gaby pour son entrain et son appétit, Nath pour sa vie en rose et ses petits délices (parfois trop secs à mon goût), Vincent pour ses lignes de poings et son expertise statistique, Pierre pour ses solos de guitares, Julie pour ses argumentations théologiques et le dernier arrivé Damien pour ses autres solos de guitares. Tout ce beau monde qui me rappelle que les pôles Bioinformatique et Applications existent bien au sein de notre équipe. Sans oublier les stagiaires que je ne peux citer par manque de place.

Je remercie aussi mes amis de St Joseph qui me rappellent que la bioinformatique n'est pas tout dans la vie, pour m'avoir supporté durant mes crises existentielles, répondu présent au téléphone et tellement d'autres choses. Par ordre alphabétique : Davy, Farid, Florian et Kevin.

Je remercie aussi mes plus anciens amis de Jussieu : C. S. Leblond et son époux J. Manry ainsi que leur chat Darwin pour leur accueil, leur bonne humeur, leur sagesse… et leur bar.

Je remercie aussi la "LBI" pour le gite, le couvert, leur gentillesse, leurs excentricités et leurs coups de sang : Anaïs, Guillaume, Naj et Sarah. Je remercie aussi la faction "Bloudi raw siz" pour les voyages, Whispering Oaks, Tristram, les pokers, l'engrenage de la "vapeur" ainsi que celui de la "toile des batailles" et la canne à sucre par ordre de morsure : Younzo, Yqnzi, Dazed&Confused, th0mdr1l aka Isorn. Si je n'ai plus grand-chose à me prouver c'est en partie grâce à vous. Mention à Soun et Philippe qui sont les seuls à me comprendre quand je parle Marvel et DC.

Je tiens aussi à remercier Naj pour avoir toujours trouvé du temps pour m'écouter, son humour, sa sagesse, son honnêteté et son amitié.

Je remercie aussi Caro pour son esprit affuté et acéré, ses verres antiques, être prête à m'accompagner pour des films indé/arti/blockbuster/navets trop souvent affligeants.

Je remercie aussi toutes les autres lignes dans le sable que j'aime tant contempler sans trop les toucher qui sont pour moi tout autant de possibilités, de décisions et d'intersections.

Je tenais à remercier tous ceux qui m'ont aidé à l'écriture de ce manuscrit, sa correction et sa mise en page. Ils se reconnaitront. S'il est présentable c'est grâce à eux.

Je remercie British American Tobacco pour son nocif soutien, ainsi que mon Mothership47 assourdissant compagnon depuis toutes ces années.

Je tenais aussi à remercier Masamoto Kishiru, son consternant art ainsi que sa science du *statu quo* m'ont servi de métronome dans ce long périple. Vil imposteur va !

A ceux que j'ai oubliés, pardonnez-moi. A ceux que j'ai pardonnés, oubliez-moi.

L'indigne que je suis allait oublier ceux qui lui ont tant sacrifié et sans qui je ne serai.

Résumé en français

Les technologies actuelles permettent d'explorer le génome entier pour identifier des variants génétiques associés à des phénotypes particuliers, notamment de maladies. C'est le rôle de la bioinformatique de répondre à cette problématique.

Dans le cadre de cette thèse, un nouvel outil logiciel a été développé qui permet de mesurer avec une bonne précision le nombre de marqueurs génétiques effectivement indépendants correspondant à un ensemble de marqueurs génotypés dans une population donnée. Cet algorithme repose sur la mesure de l'entropie de Shannon contenue au sein de ces marqueurs, ainsi que des niveaux d'information mutuelle calculés sur les paires de SNPs choisis au sein d'une fenêtre de SNPs consécutifs, dont la taille est un paramètre du programme. Il a été montré que ce nombre de marqueurs indépendants devient constant dès que la population est homogène avec une taille suffisante (N > 60 individus) et que l'on utilise une fenêtre assez grande (taille > 100 SNPs). Ce calcul peut avoir de nombreuses applications pour l'exploitation des données.

Une analyse génome-entier a été réalisée sur le photo-vieillissement. Elle a porté sur 502 femmes caucasiennes pour lesquelles un grade de photo-vieillissement a été évalué selon une technologie bien établie. Les femmes ont été génotypées sur des puces Illumina OmniOne (1M SNPs), et deux gènes ont été identifiés (*STXBP5L* et *FBX040*) associés à un SNP passant le seuil de Bonferroni, dont l'implication dans le photo-vieillissement était jusqu'alors inconnue. De plus, cette association a aussi été retrouvé dans deux autres phénotypes suggérant un mécanisme moléculaire commun possible entre le relâchement cutané et les rides. On n'observe pas de réplication au niveau du critère lentigines, la troisième composante étudiée du photo-vieillissement.

Ces travaux sont en cours de publication dans des revues scientifiques internationales à comité de lecture.

Mots clés : études d'association, entropie de Shannon, photo-vieillissement, SNP, tests multiples

Résumé en anglais

New technologies allow the exploration of the whole genome to identify genetic variants associated with various phenotypes, in particular diseases. Bioinformatics aims at helping to answer these questions.

In the context of my PhD thesis, I have first developed a new software allowing to measure with a good precision the number of really independent genetic markers present in a set of markers genotyped in a given population. This algorithm relies on the Shannon's entropy contained within these markers and on the levels of mutual information computed from the pairs of SNPs chosen in a given window of consecutive SNPs, the window size is a parameter of the program. I have shown that the number of really independent markers become stable as soon as the population is homogeneous and large enough (N > 60) and as soon as the window size is large enough (size > 100 SNPs). This computation may have several applications, in particular the diminution of the Bonferroni threshold by a factor that may reach sometimes 4, the latter having little impact in practice.

I have also completed a genome-wide association study on photo-ageing. This study was performed on 502 Caucasian women characterized by their grade of photo-ageing, as measured by a well-established technology. In this study, the women were genotyped with OmniOne Illumina chips (1M SNPs), and I have identified two genes (*STXBP5L* et *FBX040*) associated with a SNP that passes the Bonferroni threshold, whose implication in photo-ageing was not suspected until now. Interestingly, this association has been highlighted with two other phenotypes which suggest a possible common molecular mechanism between sagging and wrinkling. There was no replication for the lentigin criteria, the third component studied of photo ageing.

These studies are on the process to be published in international peer-reviewed scientific journals.

Keywords: GWAS, multitesting, photo-ageing, Shannon's entropy, SNP

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

5-HTTLPR	5-hydroxytryptamine transporter
ABCG5	ATP-binding cassette, sub-family G (WHITE), member 5
ADN	acide désoxyribonucléique
ADRB2	adrenoceptor beta 2, surface
ARN	acide ribonucléique
CNV	copy number variation
DESIR	data from epidemiological study on insuline resistance
DHFRP2	dihydrofolate reductase pseudogèene 2
DHODH	dihydrooorotate dehydrogenase
DMPK	dystrophia myotonica-proteine kinase
EIS	Estimated Independant SNPs
EM	espérance-maximisation
FBXO40	F-box protein 40
FDR	False discovery rate
GRIV	genetique de la résistance à l'infection par le VIH-1
GWAS	genomewide association study
HCG22	HLA complex group 22
HLA-C	major histocompatibility complex, class I, C
HLA-DQA1	major histocompatibility complex, class II, DQ alpha1
HLA-DQB1	major histocompatibility complex, class II, DQ beta1
HLA-DRB1	major histocompatibility complex, class II, DR beta1
HWE	hardy weinberg equilibrium
indels	insertions-délétions
LLGL4	Lethal giant larvae protein homolog 4
MAF	minor allele frequency
MC1R	Melanocortin 1 receptor
MCMC	markov chain monte carlo
Meff	M effective
NGS	next generation sequencing
OR	odds ratio
PPRX2	paired relatex homebox 2
RFLP	Restriction fragment length polymorphism

rs	reference SNP
SH3TC2	SH3 domain and tetratricopeptide repepeats 2
SNP	single nucleotide polymorphisme
STXBP5L	sintaxin binding protein 5-like
SU.VI.MAX	étude sur l'impact d'une supplémentation en vitamine et en minéraux antioxydants
TYR	tyrosinase
VIH-1	virus de l'immunodéficience humaine type 1
XP	xeroderma pigmentosum

Introduction

1. Rappel sur la génétique

a. ADN

L'Acide DésoxyriboNucléique ou ADN est le support de l'information génétique. Il constitue une matrice stable, réplicable et transmissible. Sa structure en double hélice a été découverte en 1953 par Watson et Creek. La découverte de cette molécule allait changer à jamais la compréhension du vivant comme l'avait fait Mendel à la fin du 19ème siècle. L'ADN est présent dans tous les organismes vivants, aussi bien dans les bactéries, le règne végétal que le règne animal auquel nous appartenons. Il est constitué à partir de 4 bases élémentaires appelées nucléosides ; Adénine, Thymine, Guanine et Cytosine associées à un sucre (désoxyribose) et un triphosphate, le tout formant des nucléotides communément appelés A, T, G et C (figure 1). Ces bases fonctionnent par paire une purique (Adénine ou Guanine) s'appariant avec une pyrimidique (respectivement Thymine et Cytosine) ; ce qui permet une cohésion du message génétique ainsi que la stabilité physique de la molécule ainsi constituée.

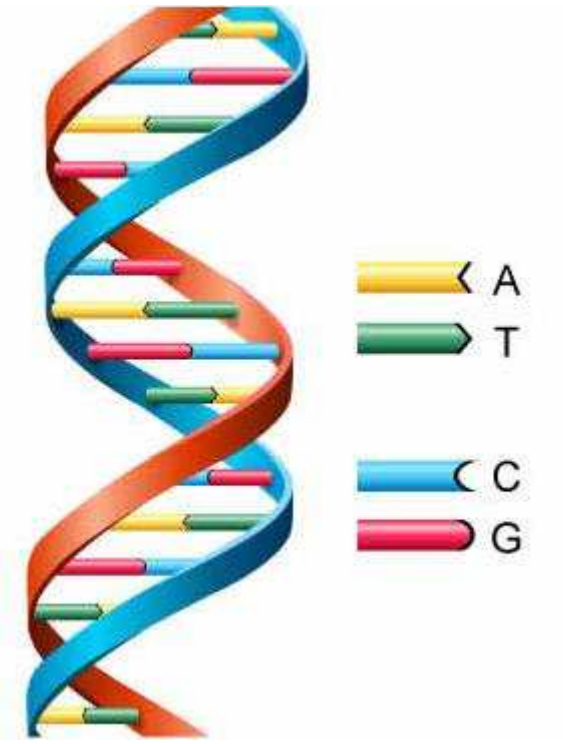

Figure 1 : Représentation schématique de la double hélice de l'ADN

Source: http://www.sciencewithmrmilstid.com/category/biological-science/

Une molécule d'ADN correspond donc à une succession de bases qui, avec son brin complémentaire, forment alors une double hélice. Cette double hélice s'associe à des protéines structurantes, appelées histones, pour s'enrouler à de multiples degrés et former une superstructure appelée chromosome. Ces différents degrés d'enroulement permettent de minimiser la place occupée par l'ADN au sein d'une cellule (figure 2). Chaque cellule contient un double exemplaire de chromosome venant de chaque parent. Chez l'homme, espèce à reproduction sexuée, on peut dénombrer 22 paires de chromosomes autosomaux ainsi qu'une paire de chromosomes sexuels. Les autosomes sont nommés par un chiffre (1 à 22) par ordre de taille décroissante, auxquels il faut ajouter les deux chromosomes sexuels X et Y. La totalité des chromosomes représente près de 3 milliards de paires de bases.

Lors de la reproduction, les chromosomes sont brassés et transmis (un chromosome de chaque paire à quelques recombinaisons près) à la descendance. L'ensemble des chromosomes est appelé génome et son étude est appelée génétique. Une partie de cette discipline se focalise sur l'ensemble des chromosomes et s'appelle génomique.

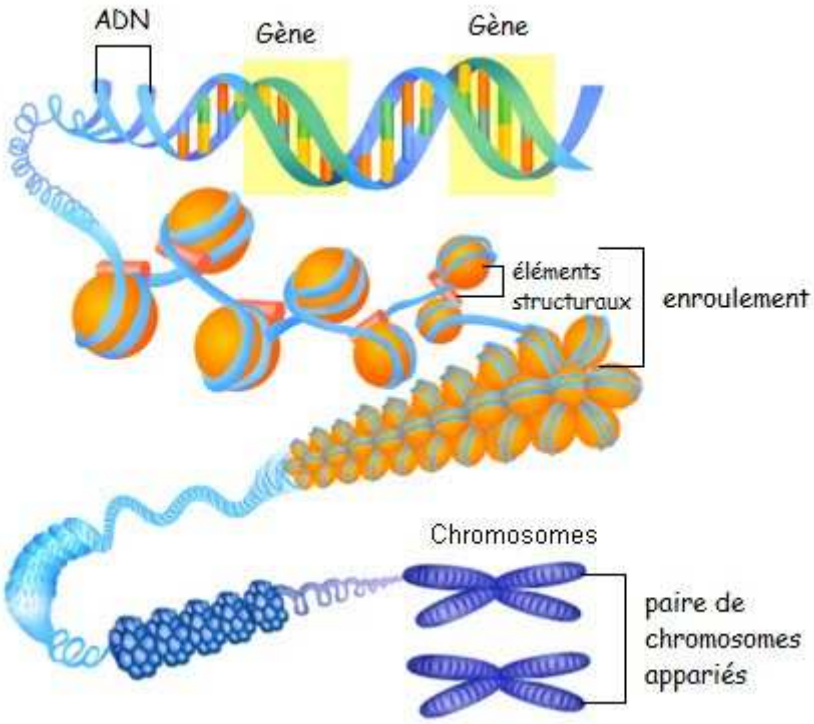

Figure 2 : Description schématique de l'ADN et de ses différents degrés d'enroulement

b. Rôle dans le vivant

L'ADN est le support de l'information génétique, présent dans chacune des cellules (à l'exception des cellules énucléées) et aussi identique à travers tout l'organisme. Le génome possède des unités fonctionnelles appelées gènes. Ces gènes sont alors transcrits en Acides RiboNucléiques (ARN) qui seront par la suite traduits en protéines (figure 3). Il est à noter que de nombreuses parties du génome ne sont pas associées à des gènes connus à ce jour, elles peuvent être des séquences régulatrices d'expression tel que des sites de fixation pour des facteurs de transcription ou bien jouer un rôle structural tel que les centromères mais plus souvent leur rôle demeure inconnu. On qualifie alors ces régions d'intergéniques.

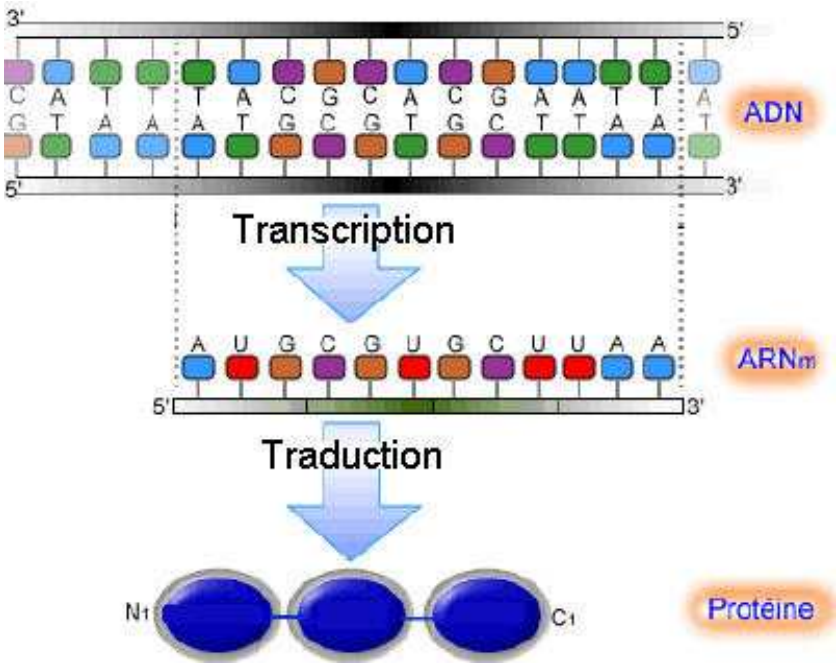

Figure 3 : Schéma des mécanismes de traduction et de transcription

L'intégrité du génome dans son enchaînement de bases nucléotidique est primordiale. Une altération du génome peut avoir des conséquences sur les unités fonctionnelles produites et donc avoir un impact sur la vie de l'organisme entier. Pour appréhender le génome, il faut rappeler que d'un homme à un autre, il existe en moyenne 3 millions de différences nucléotidiques appelés polymorphismes. Ce nombre peut paraitre important mais ramené aux trois milliards de paire de bases, il ne représente qu'un pour mille du génome [1].

c. Polymorphismes génétiques

Les polymorphismes génétiques sont des différences génétiques entre individus qui sont transmissibles d'une génération à l'autre. Ces variations s'étendent de la taille d'un chromosome à la simple variation de base nucléotidique. Les différents variants observables sont appelés allèles. Les polymorphismes ont des mécanismes de genèse différents mais ont souvent pour origine des dysfonctionnements des mécanismes de réplication du génome.

- **Polymorphismes chromosomiques**

Les polymorphismes chromosomiques sont des altérations de l'intégrité des chromosomes appelées aneuploïdies. Elles peuvent représenter des relocalisations de fragments de chromosomes telles que des translocations, inversions, fusions ou fissions (figure 4).

L'aneuploïdie peut aussi porter sur des chromosomes entiers ce qui entraînera chez l'espèce humaine des dysfonctionnements plus ou moins grave pouvant aller jusqu'à l'inviabilité de l'organisme. On peut citer dans les aneuploïdies viables, la monosomie du chromosome X appelée syndrome de Turner [2] ou bien la trisomie XXY appelée syndrome de Klinefelter [3].

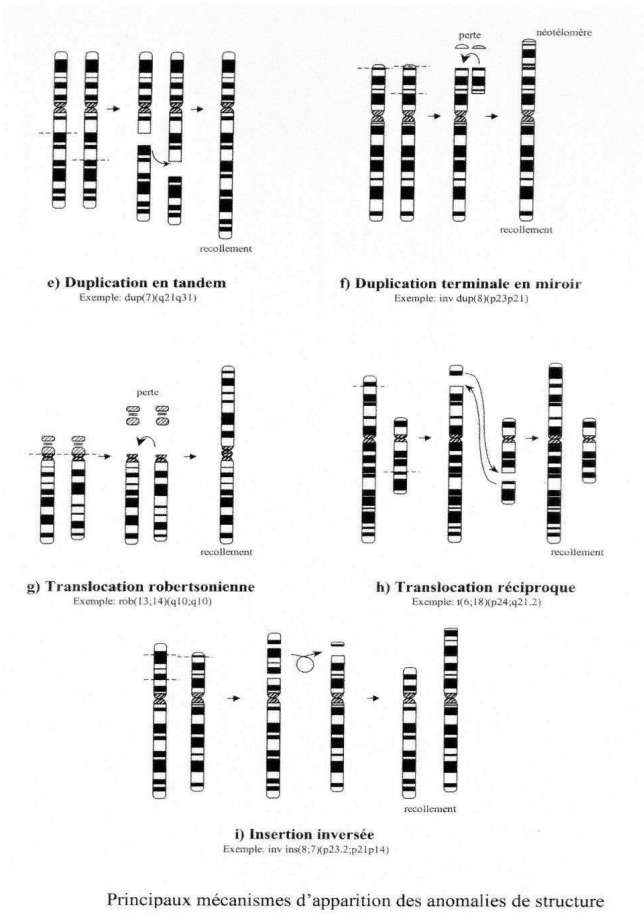

Figure 4 : Exemples d'anomalies structurales et leurs mécanismes d'apparition

Source : http://college-genetique.igh.cnrs.fr/Enseignement/genchrom/alieschrom.html

- **Séquences répétées en tandem**

Appelés satellites, minisatellites ou microsatellites en fonction de leurs tailles, ces séquences répétées en tandem correspondent à la répétition d'un motif particulier dans une séquence. Ce sont des polymorphismes multi-alléliques (plus de deux allèles observables au sein de la population). Les microsatellites sont des motifs de 1 à 5 paires de bases répétées de 2 à 50 fois pour une taille totale inférieure à 300 paires de bases ; les mini-satellites sont des motifs de 15 à 100 paires de bases répétés entre 15 à 50 fois pour une taille totale entre 1 et 5 kb ; les satellites sont de grands motifs (alpha : 171 paires de bases, beta : 68 pb) répétés en tandems entrant la plupart du temps dans des mécanismes cellulaires tels que la méiose [4].

De tels polymorphismes peuvent avoir un impact sur le bon fonctionnement d'une unité génétique comme dans le cas de la dystrophie myotonique [5] où un codon CTG est répété à plus de 37 copies dans le gène *DMPK,* perturbant ainsi la structure de la protéine.

- **Insertion-délétion**

Les insertions-délétions ou plus communément appelées indels sont des fragments nucléotidiques rajoutés ou retirés par rapport au génome de référence. Ils sont pour la plupart du temps bi-alléliques et sont notés A/AT ou bien -/T dans les bases de données. Ils s'étendent de 1pb à 1kb. En général, lorsqu'ils sont présents dans la séquence codante, ils entraînent un décalage du cadre de lecture entraînant une traduction totalement différente de l'originale. En moyenne, un être humain compte entre 192 et 280 décalages du cadre de lecture dans son génome [6].

- **Polymorphisme du nombre de copies CNV**

Les "Copy Number Variations" sont un type de polymorphisme correspondant à une large séquence d'ADN (>1kb jusqu'à plusieurs Mb) présente en un nombre variable de copies par rapport au génome de référence. Contrairement aux satellites, ils ne sont pas répétés en tandem mais à travers tout le génome. Les CNVs sont issus d'évènements d'insertion, de délétion ou de duplication, et peuvent influer sur le niveau d'expression des gènes et entraîner des pathologies [7, 8].

- **Polymorphismes mono-nucléotidique SNP**

Les "Single Nucleotide Polymorphisms" sont la plus petite forme de polymorphisme car elles n'affectent qu'une seule paire de bases. Elles constituent près de 90% des polymorphismes répertoriés. Les SNPs seront abordés plus longuement dans la partie suivante.

2. SNP

a. Présentation

Le polymorphisme mono-nucléotidique correspond à un nucléotide pour lequel on peut observer des variations au sein de la population. Hormis de rares cas, les SNPs sont des polymorphismes bi-alléliques. Leur répartition uniforme sur tout le génome et la simplicité pour les caractériser expérimentalement en font le marqueur de prédilection des chercheurs afin d'établir une cartographie dense et précise du génome (e.g. dbSNP, HapMap [9-11], le projet 1000 génomes [6]). Le nombre de SNPs connus aujourd'hui est d'environ 40 millions (source dbSNP) et ils représentent plus de 90% de la diversité génétique humaine connue. Un SNP se caractérise par sa position chromosomique, ses allèles et sa fréquence allélique mineure appelée (Minor Allele Frequency en anglais ou MAF).

Un SNP est d'abord soumis dans dbSNP en attente de validation, il a alors le statut de Submitted SNP "ss" avec un numéro unique qui lui est attribué. Puis, après validation, il acquiert le statut de Reference SNP "rs". Le SNP est alors caractérisé par ses 30 paires de bases flanquantes de part et d'autre du polymorphisme. Cette séquence peut donc être alignée sur le génome ce qui permet de déterminer la position du SNP, à savoir son chromosome et sa position sur le chromosome. Néanmoins, comme l'alignement du génome peut varier, un SNP peut être déplacé sur le génome selon les versions d'alignement utilisés (appelée "builds") au fur et à mesure que l'information du génome se précise.

Plusieurs banques de données ont vu le jour pour référencer ces SNPs et n'ont cessé de s'enrichir aussi bien en termes de SNPs génotypés mais aussi en nombre de sujets. Par exemple, le projet HapMap est passé de 1 à 3 millions de SNPs entre la phase I (2003) et la phase III (2007). De plus, HapMap a aussi augmenté le nombre de ses sujets ainsi que leur diversité, passant de 4 groupes pour un total de 270 sujets en 2003 à 11 groupes pour un total de 1301 sujets en 2007.

Le projet HapMap, quant à lui, a été développé pour étudier la structure du génome au sein de populations aux ethnicités distinctes. Le projet s'est focalisé sur des trios, à savoir un individu et ses deux parents, pour une sélection de SNPs. Le projet propose deux axes d'études majeurs à savoir : (i) les variations entre les différentes ethnies étudiées, (ii) les relations entre les SNPs au sein d'une même ethnie. Nous reviendrons par la suite sur ces relations inter-SNPs et leurs applications.

Le projet 1000 génomes propose un séquençage exhaustif du génome visant à déterminer des SNPs de fréquences faibles (MAFs inférieures à 1%) avec pour but de séquencer 2500 individus sur 28 populations différentes. Le séquençage nous offre aussi la possibilité de travailler sur des insertions-délétions. A l'heure actuelle, le projet propose une couverture de 4X sur l'ensemble du génome (un *locus* est lu en moyenne quatre fois) et de 50X dans les gènes.

b. Modèle d'équilibre d'Hardy-Weinberg

Le modèle d'équilibre d'Hardy [12]-Weinberg [13] est l'un des principes fondamentaux de la génétique des populations. Il modélise le comportement des fréquences alléliques et génotypiques pour un polymorphisme, plus particulièrement les SNPs, au sein d'une population au fil des générations sous différentes conditions. Il stipule, sous certaines hypothèses, que les fréquences alléliques et génotypiques d'un polymorphisme sont stables au sein de la population au fil des générations.

Hypothèses :

- Population de taille infinie ;
- Pangamie (union aléatoire des gamètes) et panmixie (union aléatoire des individus) :
    - Générations non chevauchantes (n'influence que les fréquences génotypiques) ;
- Absence de sélection, mutation et migration.

Modèle :

Dans le cas d'un SNP *A* bi-allélique, d'allèles a_1 et a_2, et de fréquences alléliques respectives f_{a_1} et f_{a_2} on observe que :

► les fréquences alléliques et génotypiques sont constantes au fil des générations ;

► les fréquences génotypiques suivent la distribution suivante :

$$\begin{cases} f_{a_1a_1} = f_{a_1}^{\,2} \\ f_{a_1a_2} = 2f_{a_1}f_{a_2} \\ f_{a_2a_2} = f_{a_2}^{\,2} \end{cases}$$

Lorsque les hypothèses sont respectées, on dit alors que le modèle est à l'équilibre. Toutefois, ces hypothèses peuvent ne pas être respectées : on observe alors des écarts au modèle d'équilibre d'Hardy-Weinberg. Il est à noter aussi que le modèle utilisé pour les SNPs est généralisable pour des *loci* multi-alléliques.

c. Haplotypes

Pour reprendre un terme déjà utilisé en biologie et en chimie, on parle d'un haplotype lorsque les allèles de plusieurs polymorphismes sont en position *cis*, à savoir sur le même chromosome. Il faut pour cela considérer que dans une cellule diploïde, la moitié du patrimoine génétique est transmis par la mère et l'autre moitié par le père (figure 5). On peut alors considérer les allèles des différents polymorphismes sur un même chromosome parental comme étant un haplotype. L'exemple le plus basique étant le génome mitochondrial, puisqu'il est simple chromosome et transmis par la mère, le génotypage de l'ADN mitochondrial correspond alors à la détermination de son haplotype.

Les haplotypes revêtent un intérêt biologique important de par leurs propriétés et leurs caractéristiques. Ils permettent de renseigner l'évolution des populations aussi bien humaine [14] que de tout autre organisme [15]. De plus, ils peuvent être à l'origine de dysfonctionnements non détectables en considérant de simples SNPs, comme par exemple un haplotype de plusieurs SNPs influant conjointement dans l'expression d'un gène [16] ou bien sur la fonction du gène [17].

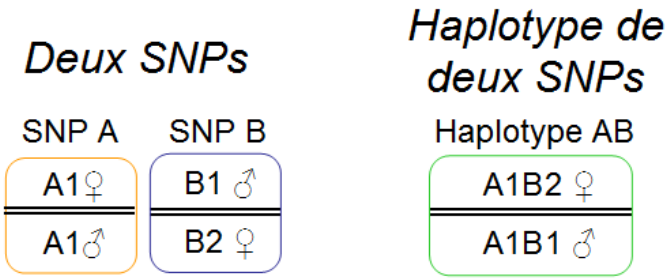

Figure 5 : Représentation schématique d'un SNP homozygote et d'un SNP hétérozygote

Dans le cas de deux SNPs, nous considérons les deux SNPs comme étant deux entités différentes et indépendantes représentées par les cadres jaune et bleu.
Dans le cadre de l'haplotype, nous considérons alors les deux SNPs comme étant une seule entité représentée par le cadre vert composée des deux SNPs précédents.
L'origine parentale est une donnée supplémentaire généralement non disponible, mise ici pour reconstituer les haplotypes.

L'haplotype permet aussi de mieux étudier le comportement des SNPs entre eux au fil des générations dans ce que nous appellerons par la suite le déséquilibre de liaison. Néanmoins, il nous faut aussi introduire une problématique de première importance concernant les haplotypes : leur reconstitution. Dans l'exemple ci-dessus, la reconstitution des haplotypes - phasage - est naturelle et sans conséquence puisque l'un des deux SNPs est homozygote et que l'erreur est sans incidence, mais la reconstitution est bien moins aisée lorsqu'il s'agit de deux SNPs hétérozygotes. Différentes méthodes seront présentées par la suite afin de reconstituer ces haplotypes. Nous y reviendrons plus loin dans le manuscrit.

d. Déséquilibre de liaison

1. Définition

Il faut introduire la notion de recombinaison qui est un phénomène qui se produit par enjambement des chromosomes homologues durant le processus de formation des gamètes qu'on appelle méiose. La probabilité qu'un évènement de recombinaison se produise entre deux *loci* chromosomiques augmente avec la distance qui les sépare. Le déséquilibre de liaison est l'association non aléatoire des allèles de deux ou plusieurs *loci* polymorphes. Lors

de la formation des gamètes, les *loci* d'un chromosome peuvent être indépendants du fait de la recombinaison et ces *loci* peuvent donc être transmis de manière indépendante (figure 6).

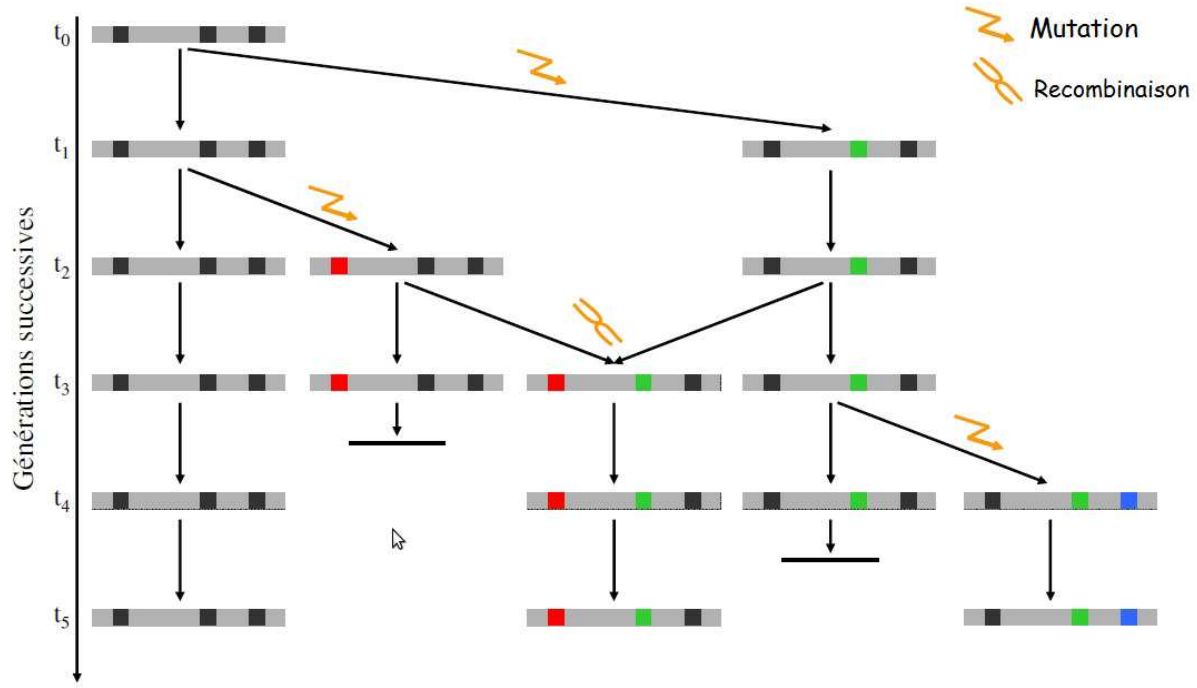

Figure 6 : Mécanismes de formation des haplotypes et du déséquilibre de liaison

Les mutations ainsi que les recombinaisons sont les seuls phénomènes à l'origine de la genèse des haplotypes.

Cependant, plus les *loci* sont proches plus la recombinaison est faible. Si nous considérons des *loci* polymorphes indépendants, les fréquences des recombinaisons d'allèles possibles sur un chromosome dans une population, correspondent alors au produit des fréquences de ces allèles, c'est l'équilibre de liaison. On peut formaliser cet équilibre entre 2 *loci* bi-alléliques :

Locus A d'allèle a_1 et a_2 de fréquences respectives f_{a_1} et f_{a_2}

Locus B d'allèle b_1 et b_2 de fréquences respectives f_{b_1} et f_{b_2}

Il existe 4 combinaisons possibles entre les allèles, les fréquences de ces combinaisons sont les suivantes :

$$\begin{cases} f_{a_1b_1} = f_{a_1} f_{b_1} \\ f_{a_1b_2} = f_{a_1} f_{b_2} \\ f_{a_2b_1} = f_{a_2} f_{b_1} \\ f_{a_2b_2} = f_{a_2} f_{b_2} \end{cases}$$

Si nous considérons des *loci* polymorphes qui ne sont pas indépendants, les combinaisons entre les allèles de ces *loci* ne se font plus au hasard. Les fréquences des combinaisons d'allèles possibles sont alors différentes du produit des fréquences alléliques, c'est le déséquilibre de liaison.

Il est important de préciser que le déséquilibre au niveau des gamètes (appelé déséquilibre gamétique) est aussi observable sur des *loci* indépendants génétiquement (chromosomes différents). Ces deux phénomènes appelés déséquilibre gamétiques et liaison génétique sont deux phénomènes différents et non réciproques. Toutefois pour l'utilisation que nous allons en faire et pour un souci de clarté nous allons assimiler les deux phénomènes en un seul que nous appellerons déséquilibre de liaison dans le reste du manuscrit.

2. Mesures

Le déséquilibre de liaison se mesure entre les fréquences haplotypiques et les fréquences alléliques. Le *D* [18] est une mesure simple du déséquilibre de liaison, pour deux SNPs *A* et *B* d'allèles a_1, a_2 et b_1, b_2 et se mesure sur les 4 haplotypes possibles :

$$\begin{cases} f_{a_1b_1} = f_{a_1} \times f_{b_1} + D \\ f_{a_1b_2} = f_{a_1} \times f_{b_2} - D \\ f_{a_2b_1} = f_{a_2} \times f_{b_1} - D \\ f_{a_2b_2} = f_{a_2} \times f_{b_2} + D \end{cases}.$$

Il est intéressant de noter que *D* est identique pour les quatre haplotypes au signe près. Lorsque les SNPs sont indépendamment distribués, alors *D* est égal à zéro.

Malgré la facilité du concept du *D*, il n'est que rarement utilisé dans la pratique car sa valeur ne renseigne pas l'importance du déséquilibre de liaison vis à vis des fréquences alléliques en jeu, rendant ainsi très difficile la comparaison entre plusieurs déséquilibres de liaison. C'est pourquoi d'autres mesures de déséquilibre de liaison normalisées ont vu le jour.

Une des mesures les plus utilisées est celle du *D*' [19] définie ci-dessous :

$$D' = \frac{D}{D_{\max}},$$

$$\text{où } D_{\max} = \min(f_{a_1} f_{b_1}, f_{a_2} f_{b_2}) \text{ lorsque } D < 0$$

$$\text{et } D_{\max} = \min(f_{a_1} f_{b_2}, f_{a_2} f_{b_1}) \text{ lorsque } D > 0.$$

Dans cette formule, *D* est normalisé avec les fréquences alléliques, il permet alors de mieux renseigner la conformation haplotypique avec nos deux SNPs. En effet, si la valeur de *D*' est égale à 1, cela signifie qu'un ou deux des haplotypes sur les quatre théoriquement

possibles, sont manquants. Cela permet de repérer soit des événements de mutations récents soit un déséquilibre de liaison total entre les deux SNPs.

La mesure la plus utilisée est celle du R^2 [20] définie ci-dessous :

$$R^2 = \frac{D^2}{f_{a1} f_{a2} f_{b1} f_{b2}}.$$

Cette fois ci, *D* est normalisé avec toutes les fréquences alléliques en jeu. Il permet alors d'avoir une information capitale : le déséquilibre de liaison total $R^2=1$. Dans cet état, on peut donc déduire que :

- les deux SNPs ont les mêmes fréquences alléliques ;
- il n'y a plus que deux conformations haplotypiques observées.

Dans ce cas précis, le SNP *B* est le reflet du SNP *A* et ils contiennent la même information d'un point de vue génomique, bien qu'ils puissent avoir des effets différents d'un point de vue fonctionnel. On dit alors que le SNP *A* tague le SNP *B* et vice-versa, l'un des deux SNPs pouvant servir de TagSNP. On considère que deux SNPs sont en déséquilibre de liaison lorsque leur R^2 est supérieur à 0,80 (figure 7).

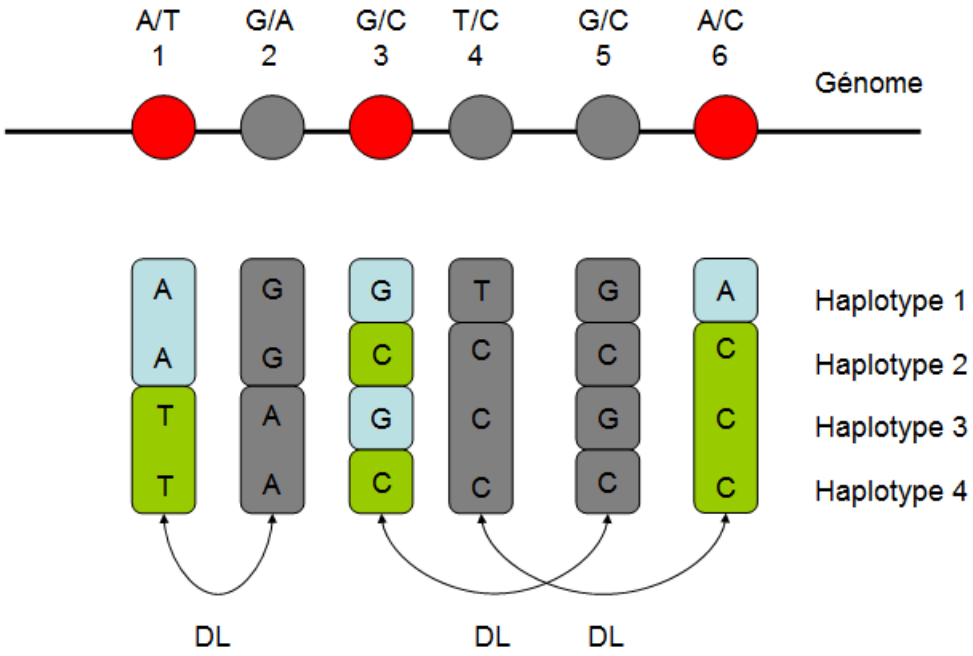

Figure 7 : Représentation de Tag SNPs

Les SNPs rouges tagguent les SNPs gris. L'allèle d'un TagSNP permet de connaître l'allèle du SNP taggué. Les 3 TagSNPs (1,3 et 6) permettent de renseigner l'ensemble de l'haplotype.

Le déséquilibre de liaison diminue au fil des générations en fonction de la distance génétique en centiMorgans selon la formule :

$$D_{AB}(t+1) = (1-c)D_{AB}(t) \ ;$$

avec *t*, le temps en génération, et *c*, le taux de recombinaison entre les deux *loci*[21].

Ces trois mesures *D*, *D'* et R^2 sont les mesures les plus utilisées pour leur simplicité aussi bien au point de vue du concept mais aussi des informations qu'elles nous permettent d'obtenir. Néanmoins, elles présentent toutes les trois le même écueil, au sens où elles ne sont pas (ou bien très difficilement) généralisables sur de multiples *loci* (3 et plus). D'autres mesures ont été développées [22, 23], nous verrons par la suite une autre mesure de déséquilibre de liaison que nous utiliserons pour mesurer le déséquilibre de liaison sur plusieurs *loci*.

e. Reconstruction des haplotypes

1. Problématique

Les haplotypes sont donc la succession d'allèles de différents *loci* portés par le même brin d'un chromosome, en position *cis*. Néanmoins, ces informations ne sont pas disponibles ; en effet, les technologies de séquençage et de génotypage ne fournissent pour l'heure que des informations sur les génotypes. C'est à partir de cette base de génotypes que l'on peut déterminer la "phase" des deux SNPs, c'est à dire les haplotypes présents dans cette paire de SNPs.

Dans le cas où deux SNPs sont présents, il est assez simple de reconstruire la phase dans la plupart des conformations. Néanmoins, dans le cas d'un haplotype composé de deux SNPs hétérozygotes l'incertitude réside dans la phase (tableau 1). Dès lors, il faudra utiliser des méthodes statistiques basées sur le déséquilibre de liaison pour déterminer la probabilité des haplotypes dans notre cas. A noter aussi que pour *n loci* bi-alléliques, nous avons donc 2^n haplotypes possibles même si dans la pratique le déséquilibre de liaison réduit le nombre d'haplotypes observés. Il devient alors nécessaire de recourir à des logiciels bioinformatiques pour prédire les haplotypes dans des populations de sujets non apparentés par exemple dans le tableau 1, le double hétérozygote. Toutefois, dans les données génomiques familiales, il est plus facile de reconstituer les haplotypes sans nécessairement avoir recours à une grande logistique informatique et bioinformatique.

SNP *A* / SNP *B*	$\frac{a_1}{a_1}$	$\frac{a_1}{a_2}$	$\frac{a_2}{a_2}$
$\frac{b_1}{b_1}$	$\frac{a_1b_1}{a_1b_1}$	$\frac{a_1b_1}{a_2b_1}$	$\frac{a_2b_1}{a_2b_1}$
$\frac{b_1}{b_2}$	$\frac{a_1b_1}{a_1b_2}$	$\frac{a_1b_2}{a_2b_1}$ *ou* $\frac{a_1b_1}{a_2b_2}$	$\frac{a_2b_1}{a_2b_2}$
$\frac{b_2}{b_2}$	$\frac{a_1b_2}{a_1b_2}$	$\frac{a_1b_2}{a_2b_2}$	$\frac{a_2b_2}{a_2b_2}$

Tableau 1 : Haplotypes possibles pour deux SNPs.

Cases vertes : génotypes ; Cases blanches : paires d'haplotypes certains

Case rouge : paires d'haplotypes possibles et incertains

2. Méthodes d'inférence des haplotypes

Les techniques de génotypage ou de séquençage ne permettant pas de distinguer le chromosome parental où se trouve l'allèle muté, on ne dispose que de génotypes. Ainsi pour deux SNPs voisins sur un même chromosome, le premier ayant deux allèles A ou C, le deuxième G et T, on sait qu'un individu pourra être AA, AC, ou CC sur le premier et GG, GT, TT sur le deuxième SNP. Mais on ignore si un individu hétérozygote pour ces deux SNPs aura pour combinaison haplotypique AG/CT ou AT/GC. Divers logiciels reposant sur plusieurs modèles, des plus simples aux plus complexes, permettent de reconstruire les haplotypes, c'est-à-dire de "phaser les individus" :

- **les méthodes combinatoires** reposant sur le principe de tester toutes les combinaisons d'haplotypes possibles et de les discriminer via un critère de parcimonie ou bien de phylogénie [24],
- **les méthodes d'inférence statistique** reposant sur la vraisemblance ou encore sur des algorithmes bayésiens ou pseudo-bayésiens.

Les méthodes basées sur la vraisemblance utilisent un algorithme d'Espérance-Maximisation (EM) [25] qui est un algorithme itératif permettant d'évaluer les haplotypes les plus vraisemblables correspondant aux observations.

L'algorithme se décompose ainsi :

Soit k haplotypes possibles dans la population de fréquences respectives f_k :

étape 0 - randomisation : attribution aléatoire d'une valeur pour chaque fréquence $f_k(0)$;

étape 1 - espérance : calcul de l'espérance de la vraisemblance des observations (probabilité d'observer les génotypes des individus avec f_k) ;

étape 2 - maximisation : estimation du maximum de vraisemblance pour chacune des fréquences $f_k(1)$ (réajustement des f_k à partir des observations et leur probabilités) ;

étape 3 - évaluation : si l'on considère que l'algorithme a convergé ($f_k(t) \cong f_k(t-1)$) alors on s'arrête, sinon on retourne à l'étape 1 avec les nouvelles valeurs $f_k(1)$.

Quant aux méthodes basées sur des algorithmes bayésiens ou pseudo-bayésiens, elles se servent à la fois de l'information a priori des fréquences haplotypiques et de l'information des génotypes pour calculer la distribution a posteriori des haplotypes, sachant les génotypes observés. Le modèle de distribution des haplotypes utilise le modèle de coalescence et prend en compte les recombinaisons lors de la méiose. Un échantillonneur de Gibbs (algorithme MCMC) est utilisé pour approximer cette distribution à partir des génotypes observés.

La première méthode de phasage basée sur l'algorithme d'espérance-maximisation a été introduite par Excoffier et Slatkin en 1995 [26]. Les méthodes utilisant les algorithmes bayésiens ou pseudo bayésiens ont été introduites en 2001 avec le logiciel PHASE [27] et plus récemment Shape It [28, 29] permettant enfin de phaser un chromosome entier. D'autres méthodes ont aussi été développées en prenant en compte les haplotypes fondateurs permettant de raffiner et d'accélérer la qualité de l'imputation.

Ceci permet de mettre en évidence les progrès de la bioinformatique, aussi bien en termes de précision mais aussi de rapidité, afin de répondre à des problématiques croissantes. L'haplotypage constitue un pan entier de la génomique comme peuvent en témoigner les différentes revues qui lui sont consacrées aussi bien pour la reconstruction des haplotypes [30-33] ou l'analyse des haplotypes [34].

3. Importance des marqueurs génétiques dans l'étude des maladies

L'intervention de facteurs héréditaires dans les maladies a été observée depuis des siècles, mais la description objective d'une variation génétique causale de phénotypes particuliers remonte aux années 1950 avec la découverte par Lejeune [35] de la trisomie du chromosome 21 associée au phénotype de "mongolisme". Avec les progrès de la génétique moléculaire dans les années 1980, il a été possible d'identifier des *loci* génétiques précis associés à des maladies familiales, ces maladies ont été qualifiées de monogéniques car dues à la défaillance d'un seul gène.

Au fur et à mesure que les techniques de biologie moléculaires ont progressé, il a été possible de couvrir le génome par des marqueurs génétiques de plus en plus fins. Les premiers marqueurs furent les polymorphismes de longueurs des fragments (RFLP) puis les mini satellites et les micro satellites. Les derniers marqueurs a avoir été exploités sont les SNPs sur lesquels nous allons nous attarder.

En fonction de la relation entre la population d'étude et le phénotype, les études peuvent donc se diviser en deux catégories études de liaison qui se focalisent la transmission des gènes et celui du phénotype au sain d'individus et les études d'association qui se focalisent la corrélation entre nos marqueurs et celle SNPs. De plus les études se divisent aussi en deux catégories selon la couverture du génome étudiée, à savoir la couverture de quelques gènes, appelée études gènes candidats, et la couverture intégrale du génome, appelée génome entier. Les études les plus populaires en génomique aujourd'hui sont les études d'association génome entier appelées GWAS (Genome Wide Association Study) en anglais.

Après avoir décrit les différents types d'études possibles (familiale : étude de liaison, ou sur des individus non apparentés : études d'association), et le type de récolte d'informations sur notre cohorte (transversale ou longitudinale), nous aborderons alors la question de la région du génome à étudier.

a. Études de liaison

Les études de liaisons sont les premières études à avoir vu le jour. Elles étudient la co-ségrégation d'un phénotype et d'un génpotype au sein de familles en d'autres termes en

cherchant un polymorphisme qui se transmette de la même façon que le phénotype (figure 8). Ce type d'étude permet d'identifier des facteurs génétiques liés à des traits monogéniques tels que la mucoviscidose [36, 37] ou la chorée de Huntington [38-40].

Les études de liaisons sont particulièrement efficaces pour trouver des traits dits mendéliens, mais sont soumises à quelques limitations :

- les arbres généalogiques sont parfois incomplets ou insuffisants ou encore on peut ne disposer que d'un seul membre de la famille touchée ;
- efficacité limitée dans les maladies multifactorielles où chaque facteur n'explique qu'une fraction du génotype, ou bien lorsque les traits étudiées ont des composantes externes.

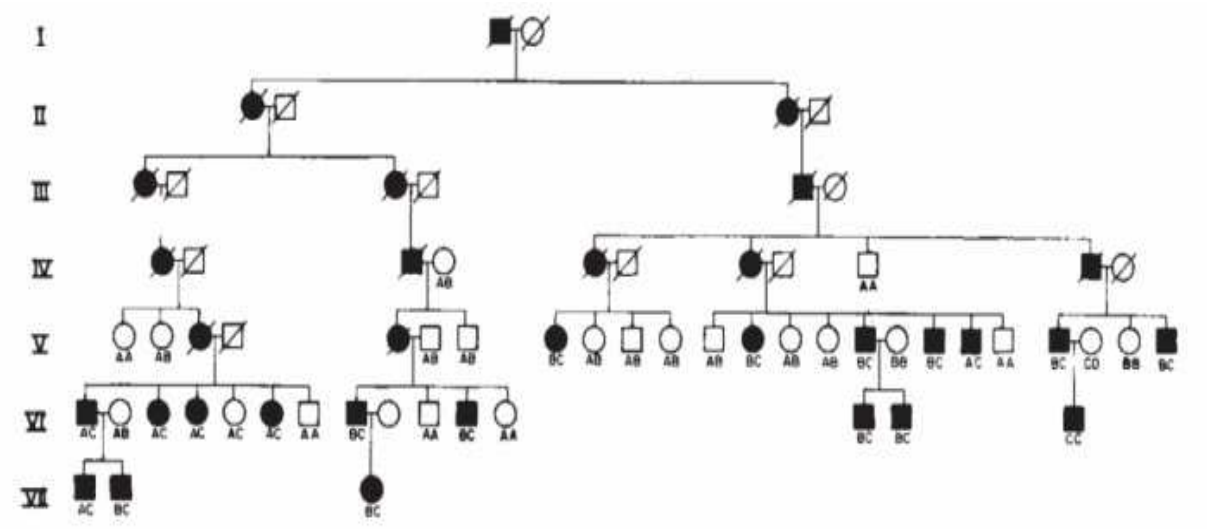

Figure 8 : Exemple de pédigree tiré d'une étude sur la chorée d'Huntington

Cas d'une famille vénezuellienne [40]*, les ronds représentant les femmes et les carrés les hommes,les blancs représentant les individus non atteints et les noirs les individus atteints,les barrés représentant les individus décédés,les génotypes sont sous les individus.*

Ces études ont été les premières à avoir vu le jour et elles restent encore d'actualité comme en attestent les récentes études pour l'autisme [41, 42] ou bien pour la schizophrénie [43, 44].

b. Études d'association

Les études d'association comparent la répartition des allèles en fonction du trait étudié entre les individus qui portent le trait et ceux qui ne le portent pas. Plus la répartition est différente, plus le SNP est susceptible d'être impliqué avec le trait étudié. Elles sont réalisables aussi bien sur des individus apparentés que sur des non apparentés mais nous n'allons décrire uniquement les études d'association sur des individus non apparentés. Dans ce cas, on cherchera alors à disposer d'une population homogène et non apparentée.

Ces études se sous-divisent en deux catégories : les études transversales et les études longitudinales selon le type de données et la logistique disponible en termes de recueil des données.

Les études transversales se focalisent sur un moment donné et s'apparentent à un "cliché" de la situation. On recherche une corrélation entre le trait et une variable explicative (figure 9), dans notre cas un SNP, se traduisant par une différence de répartition des variables explicatives (génétiques) vis à vis des variables à expliquer (trait phénotypique).

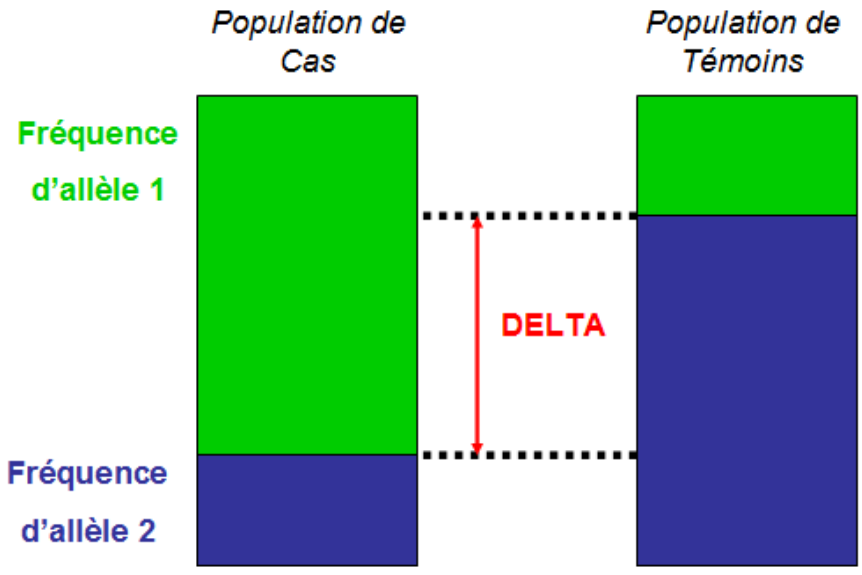

Figure 9 : Représentation schématique d'une étude d'association dans une étude transversale 'Cas/Témoins.

Delta représentant l'écart de fréquences observé entre les Cas et les Témoins.

Les études longitudinales suivent l'évolution d'un trait sur une période donnée. Une étude longitudinale s'apparente à un "film" par son suivi régulier des patients afin d'avoir le plus "d'images" possibles pour obtenir une définition plus précise de l'évolution du trait au cours du temps.

Les courbes de survie (figure 10) permettent de visualiser la survenue d'un évènement (apparition d'un trait phénotypique, par exemple la mort) au cours du temps pour les différents groupes étudiés à risque supposés différents. Dans le cas d'études génomiques, les différents groupes étudiés sont alors les génotypes.

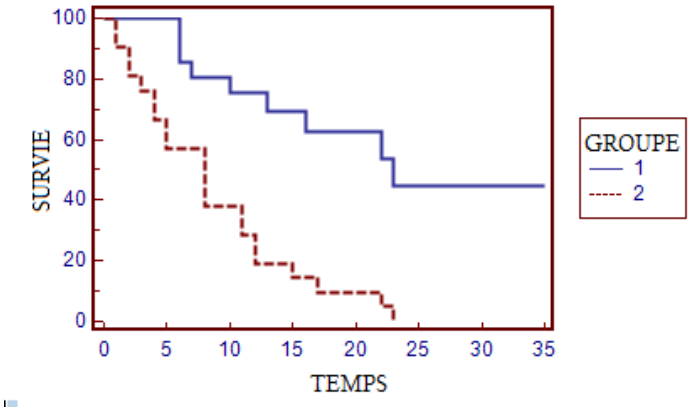

Figure 10 : Représentation d'une courbe de survie

Courbe de survie avec en abscisse le temps et en ordonné le pourcentage de survie au sein des groupes

Les études longitudinales nous permettent de suivre l'évolution de groupes de sujets jusqu'à ce qu'ils atteignent l'état étudié (apparition du trait). Cela suppose une logistique de collecte de données importante car il faut que les sujets soient suivis régulièrement afin d'avoir une estimation précise de la survenue de l'évènement, en particulier l'apparition des maladies. De par leur construction, ce type d'études permet de mettre en évidence des gènes associés à des maladies longues telles que des cancers [45-47] ou des maladies coronaires [48, 49] plus précisément dans la survenue d'évènements (mort, rechute, arrivée d'un accident) ou bien de réponse aux traitements (survie, temps avant rechute) une fois que la maladie a été diagnostiquée.

Les représentations schématiques (figures 9 et 10), montrent que l'on peut visuellement discriminer les deux groupes et supposer qu'un des groupes est effectivement

corrélé à l'état étudié ou à une survenue du trait étudié. Nous décrirons plus loin les moyens permettant de quantifier ces différences observées.

c. Études gènes candidats

Si le trait phénotypique est bien renseigné et que nous avons déjà des informations concernant les mécanismes moléculaires ou génétiques entrant en jeu, nous pouvons nous restreindre à l'étude d'une où des régions bien définies, le plus souvent centrées sur des gènes (expliquant ainsi le nom de ce type d'analyses). Dans ce type d'analyse, on se focalise donc sur une région prédéfinie par des a priori biologiques, ce qui nous permet de concentrer nos moyens et donc d'obtenir une cartographie très fine via le séquençage de la région.

Dans ce cas de figure, nous pouvons travailler sur une liste de polymorphismes génomiques caractérisés de manière exhaustive sur la région étudiée, tels que les insertion/délétions, les séquences répétées en tandem, les SNPs de fréquence commune ou rare (MAF <1%) voire des singletons (un seul porteur de l'allèle au sein de toute la population). Enfin, cette cartographie fine peut permettre d'identifier des polymorphismes qui n'étaient pas encore connus.

Malgré les avantages indéniables de l'approche gène candidat au niveau des polymorphismes caractérisés, elle souffre d'un écueil notable : sa nécessité de connaissances sur le rôle possible du gène a priori. Les études gènes candidats permettent en revanche d'approfondir des connaissances déjà acquises et de valider ou non des hypothèses préalables, mais rarement de découvrir des mécanismes *ex nihilo.* Pour ces études, il est possible d'utiliser le séquençage direct ou des puces de génotypage à façon (dont le principe va être décrit juste après) qui impliquent d'avoir une connaissance préalable des polymorphismes de la région.

d. Études génome entier

Lorsque l'on cherche à découvrir de nouveaux mécanismes biologiques, il est important de partir sans hypothèse génétique *a priori* sur les traits étudiés. Les progrès de la biochimie ont permis de franchir cette barrière et de génotyper (caractériser le génotype d'un sujet) le génome entier par le biais de puces de génotypage.

Il y a principalement deux sociétés qui proposent actuellement des puces de génotypage reposant sur deux méthodes biochimiques différentes mais aussi sur deux approches différentes dans le choix de SNPs. Affymetrix a choisi ses SNPs à un intervalle régulier sur le génome, faisant abstraction du déséquilibre de liaison. Illumina a choisi ses

SNPs en sélectionnant des TagSNPs afin de maximiser l'information du génome en utilisant le moins de marqueurs possibles. Les premières puces Affymetrix renseignaient environ une centaine de milliers de SNPs répartis sur tout le génome. Elles contiennent aujourd'hui jusqu'à 2,5 millions de SNPs, ainsi que des polymorphismes de types CNVs.

Les puces permettent de cartographier finement le génome entier et d'identifier des régions d'intérêt. Néanmoins, de par le choix des polymorphismes présents sur la puce, seule une fraction (même grande) du génome est renseignée, étant considérée comme intéressante par les sociétés fabricantes. Il faut aussi ajouter que les SNPs ayant une MAF de fréquence faible sont difficilement génotypés et que les indels sont aussi absents des puces. Ces puces ne font encore une fois que repousser la limite de l'*ex nihilo* mais ne sont pas exhaustives.

Les progrès en terme de densité ainsi que la réduction de leur coût ont tout de même fait des puces de génotypage l'outil standard d'identification de nouveaux gènes impliqués dans des maladies humaines comme en témoigne la base de données GWAS catalog (www.genome.gov/gwastudies)[50] répertoriant plus de 1200 publications référençant près de 6200 SNPs significativement associés. De plus ceci ne concerne que les études d'association, il est possible aussi de mener une étude de liaison à travers le génome entier.

Outre les questions du type d'études et de phénotypes observées, les études génome entier ont eu pour conséquence l'accroissement de la quantité des données à traiter. Les centaines de milliers ou les millions de SNPs ne peuvent être analysés manuellement et nécessitent l'intervention de bioinformaticiens et biostatisticiens pour prendre en charge toute la logistique des données informatiques mais aussi pour l'analyse de ces mêmes données.

4. Analyse d'association sur génome entier

Dans cette partie, je vais décrire en détail les techniques permettant l'analyse d'association sur génome entier qui a constitué la majeure partie de mon travail de thèse, ce travail est le cœur du savoir-faire de notre équipe dans le but de trouver de nouvelles pistes diagnostiques et thérapeutiques et de confirmer des hypothèses. Tout d'abord, je vais présenter le génotypage qui fournit les données pour l'analyse, des méthodes de contrôle de qualité du génotypage, les méthodes de comparaison des distributions et de contrôle de qualité de l'analyse et enfin les méthodes de recherche post-association.

a. Génotypage

1. Puces de génotypage

Les progrès de la biochimie associés aux cartographies de SNPs de plus en plus renseignées ont permis de tirer profit des redondances au sein du génome en identifiant des SNPs représentatifs de tous les autres, les TagSNPs. Nous pouvons alors par le biais du génotypage d'un nombre restreint de TagSNPs déterminer le génotype d'un plus grand nombre de SNPs connus. La quantité croissante d'information observable couplée à la réduction des coûts du génotypage ont permis l'explosion du nombre de GWAS avec près d'un millier d'études publiées en seulement quelques années.

La technologie Illumina, utilisée au sein du laboratoire, repose sur l'hybridation d'ADN génomique (après amplification et fragmentation) avec des oligomères de 50 bases couplées à des billes. Il s'ensuit alors une extension spécifique de l'oligomère avec une base marquée. Chaque bille est spécifique d'un SNP et chaque base nucléotidique possède son fluorophore spécifique ce qui permet de relever des intensités de fluorescences différentes selon les deux allèles possibles pour chaque SNP. Pour chaque SNP, les intensités obtenues pour l'ensemble des échantillons de la population sont traitées par un programme qui va inférer les génotypes de chaque individu (figure 11).

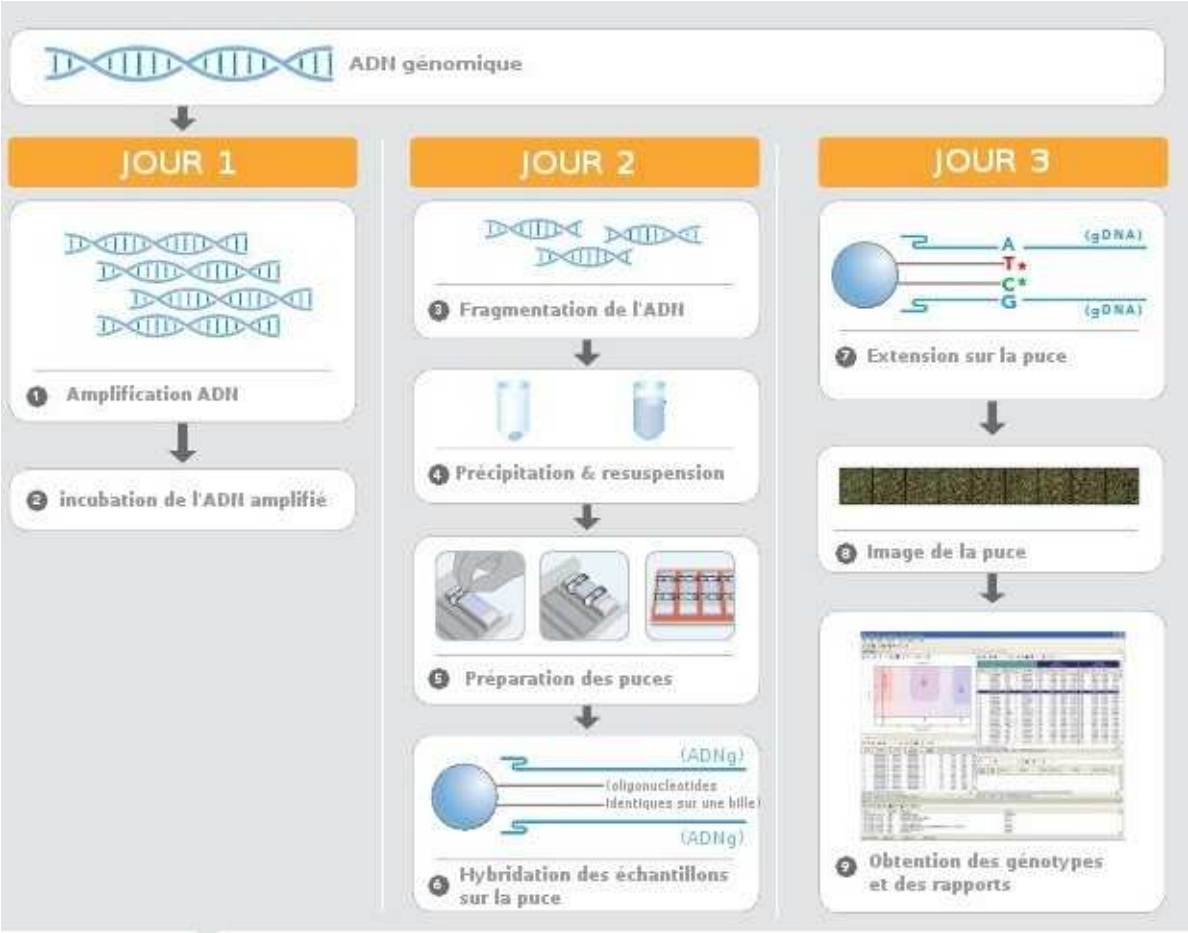

Figure 11 : Présentation du protocole expérimental de la technologie Illumina

2. Inférence des génotypes

Bien que le concept soit simple en apparence, une couleur pour chaque génotype, l'inférence des génotypes se révèle en pratique plus complexe. En effet, de nombreux biais expérimentaux peuvent influencer les intensités de fluorescence et troubler les mesures. Pour y remédier Illumina propose son propre logiciel d'inférence de génotype afin de normaliser les intensités sur la population et de déterminer quels génotypes seront attribués à quelles intensités. L'algorithme d'inférence population reste de l'ordre du savoir-faire secret de la compagnie Illumina. Toutefois, il peut arriver que l'algorithme n'arrive pas à décider, et même parfois, que l'algorithme se trompe et que certains SNPs soient mal inférés.

C'est pourquoi le fabricant propose deux scores de contrôle qualité appelé "call rate" qui correspond à la fréquence de SNPs inférés par individus et "call frequency" qui correspond à la fréquence de SNPs inférés par locus.

b. Contrôle de qualité du génotypage

Plusieurs types de scores permettent de filtrer la qualité de l'inférence des SNPs :

- **Fréquence allélique mineure MAF** : cette fréquence sert de filtre pour le contrôle de la qualité car pour un algorithme d'inférence il est très difficile de génotyper un SNP de fréquence faible. En effet, dans l'espace de représentation des intensités lors de l'inférence des génotypes, les fréquences faibles produisent de petits ilots d'un ou deux individus pour les homozygotes portant l'allèle mineur, de ce fait ils risquent d'être assimilés à tort au groupe le plus proche, à savoir les hétérozygotes. Plusieurs seuils peuvent être utilisés selon les effectifs de la population à l'étude, mais les valeurs sont en général fixées entre 1% ou 5%. De plus, il est possible de vérifier à l'œil nu les groupes d'intensités et donc de valider visuellement la qualité du génotypage. C'est ce que nous avons fait dans une des études de notre équipe sur les SNPs fréquences faibles impliqués dans la progression vers le SIDA [51] ;
- **Équilibre d'Hardy-Weinberg HWE** : L'équilibre d'Hardy-Weinberg est aussi utilisé comme garant de la bonne tenue du génotypage. Dans la mesure du possible, le filtre est appliqué dans la population témoin afin d'être sûr que des SNPs liés à des écarts au modèle d'Hardy-Weinberg ne viennent perturber l'analyse. Il est aussi utilisé pour vérifier qu'il n'y ait pas d'erreurs expérimentales.

Bien qu'aucun seuil universel n'ait été établi, un consensus s'est assemblé autour d'un seuil de p-value (notion définie par la suite) de 5*10-3 pour le test d'adéquation à la loi d'Hardy-Weinberg en dessous duquel le SNP est écarté ;

- **Données manquantes :** A partir d'un certain seuil de données manquantes (génotype pour lequel l'algorithme d'inférence n'a pu trouver de solution acceptable selon ses propres critères), on peut considérer qu'il y a eu un souci sur la puce au niveau de plusieurs SNPs ou bien une erreur expérimentale sur plusieurs individus. Lorsque qu'un souci se produit sur une puce, l'individu génotypé présente une proportion trop importante de données manquantes et sera écarté lors du filtrage des données manquantes par individu. Lorsqu'il y a un défaut dans les puces, alors certains *loci* étudiés présenteront une proportion trop importante de données manquantes. Le *locus* sera donc écarté lors du filtrage par données manquantes par SNP.

Bien qu'aucun seuil universel n'ait été établi, un consensus s'est assemblé autour de 2% comme seuil de rejet pour les données manquantes par individu et pour les SNPs génotypés.

c. Association entre un SNP et un phénotype

Une fois les différentes étapes du contrôle de qualité effectuées, il est ensuite possible d'étudier l'association d'un SNP avec phénotype étudié. Afin de déterminer si notre SNP est associé au phénotype étudié, nous allons déterminer si la répartition allélique ou génotypique de notre SNP est corrélée à la dichotomie de nos populations cas et témoins. Plusieurs modèles génétiques peuvent être utilisés afin de définir la répartition des allèles ou des génotypes et ainsi identifier une association potentielle.

Par la suite et hors mention contraire, nous considérerons que nous somme dans le cadre d'une dichotomie cas/témoins avec un phénotype binaire de type (atteint/non atteint). Nous considérerons un SNP A d'allèle a_1 et a_2 et de génotypes a_1a_1 (appelés homozygote a_1a_1), a_1a_2 (appelés hétérozygote) et l'homozygote a_2. Sauf mention contraire, nous considérerons que l'allèle a1 causal du phénotype (atteint).

1. Répartition allélique

Sans se baser sur un modèle génétique, nous pouvons tester si l'allèle a1 est associé au phénotype sans prendre en compte les individus mais simplement les allèles en présence. Nous pouvons alors diviser nos populations cas et témoin par leurs comptes d'allèle a_1 (N_{a1} = 2 * $N_{homozygote\ a1}$ + $N_{hétérozygote}$) que l'on opposera a leur compte d'allèle a_2 (N_{a2} = $N_{hétérozygote}$ + 2 * $N_{homozygote\ a2}$).

2. Répartition génotypique et modèle génétique

Nous pouvons diviser nos populations de cas et de contrôles selon leurs génotypes en suivant des modèles génétiques qui sont des hypothèses de fonctionnement des deux allèles a_1 et a_2 et de leur impact sur le phénotype. Ils permettent de partitionner les populations en fonction des génotypes.

- modèle additif : On suppose que porter une seule copie de l'allèle a1 a un effet moitié moindre sur le phénotype que de porter deux copies. Le nombre d'allèle a_1 portés produit donc un effet dose sur le phénotype. Nous sortons alors du cadre d'un phénotype strictement binaire. Nous pouvons donc répartir nos populations de cas et de témoins selon leur nombre de copie de l'allèle a_1 tout en gardant à l'esprit l'effet dose de celui-ci.

- modèle dominant : On suppose que le simple fait de porter une copie de l'allèle a_1 suffit à entrainer le phénotype (atteint). Nous pouvons donc diviser nos populations de cas et de témoins selon le fait qu'ils portent au moins une copie de l'allèle a_1 (homozygote a_1 et hétérozygote) ou non (homozygote a_2).

- modèle récessif : On suppose alors qu'il faut porter deux copies de l'allèle $a1$ pour entrainer le phénotype (atteint). Nous pouvons donc diviser nos populations cas et témoins selon le fait qu'ils portent deux copies de l'allèle a1 (homozygote a_1) ou non (hétérozygote et homozygote a_2).

- modèle codominant : On suppose cette fois ci que les deux allèles (a_1 et a_2) possèdent chacun leur propre effet sur le phénotype. Les deux homozygotes ont donc leurs phénotypes propres et l'hétérozygote a un phénotype différent des deux précédents. Nous sortons alors du cadre du phénotype binaire. Il n'inclut pas d'effet dose notre phénotype est donc divisé en trois classes sans hiérarchisation.

- répartition génotypique : Sans supposition d'un modèle génétique a priori, nous considérons chaque génotype comme étant une modalité de notre répartition de SNPs.

d. Test d'hypothèses

- **Rappel de la problématique**

Notre problématique est de déterminer si pour un SNP donné la répartition de nos allèles ou de nos génotypes est différente selon notre phénotype (atteint/non atteint). Le principe du test statistique consiste à supposer que la distribution n'a pas d'influence sur le phénotype et que la probabilité d'observer la distribution est donc le simple jeu du "hasard".

- **Hypothèses**

Un test d'hypothèse revient donc à opposer deux hypothèses H_0 contre H_1, avec un critère de décision permettant de choisir entre ces deux hypothèses.

H_0 hypothèse nulle

Elle constitue l'hypothèse de prudence si les résultats du test ne sont pas concluants. Le fait de ne pas rejeter H_0, ne permet pas de prouver qu'elle soit vraie, elle est "acceptée par défaut" car les résultats n'ont pas été suffisants contre elle, elle est simplement présumée vraie.

H_1 hypothèse alternative

Elle constitue l'autre hypothèse. On dit que l'on teste H_0 contre H_1. Différentes formulations sont possibles pour H_1 pour mieux préciser cette différence, dont je vais vous présenter deux cas utilisé en biologie :

Dans ce qui suit, θ est un paramètre relatif à une population, donc inconnu et θ_0 est le paramètre observé. Deux situations sont possibles : lorsque l'on n'a pas d'idée privilégiée d'un sens de la différence le test est bilatéral sinon unilatéral.

- Test bilatéral

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

Dans ce cas on teste la différence au sens stricte du terme, comme par exemple un test de conformité de la taille moyenne d'un échantillon par rapport à celle de la population.

- Test unilatéral

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta > \theta_0$$

Dans ce cas, on s'attend un effet supérieur, comme par exemple un médicament qui aurait un effet supérieur comparé à celui du placebo.

- **Niveau et puissance**

On détermine alors la loi de distribution de la variable aléatoire sous H_0 qui représente alors toutes les valeurs possibles de notre distribution. Cette distribution comprend alors deux zones : la zone d'acceptation, où H_0 est admise (ou non rejetée) et la zone de rejet ou région critique, où H_0 est rejetée au profit de l'hypothèse alternative H_1 en fonction d'un seuil de significativité.

Toutefois, le hasard de l'échantillonnage peut fausser les conclusions et quatre cas sont observables (Tableau 2).

		Réalité	
		H_0 vraie	H_0 fausse
Décision	Non Rejet de H_0	ok	erreur de deuxième espèce
	Rejet de H_0	erreur de première espèce	ok

Tableau 2 : Tableau de décision et erreurs possibles

Si l'hypothèse retenue correspond à la réalité alors la conclusion est correcte, mais il y a deux cas où l'on conclut à tort :

- L'erreur consistant à rejeter à tort l'hypothèse vraie H_0 est appelé erreur de 1ère espèce, l'affirmation de quelque chose qui ne l'est pas, dans notre cas un faux positif et sa probabilité de survenue appelée risque α ;
- L'erreur consistant à accepter à tort l'hypothèse fausse H_0 est appelée erreur de 2ème espèce, le fait de ne pas déceler ce qui est présent, dans notre cas un faux négatif et sa probabilité de survenue appelée risque β.

Le risque de 2ème espèce permet aussi de calculer la puissance d'un test ($P = 1 - \beta$) qui correspond à la capacité d'un test à différencier la distribution de l'échantillon de celle de la population. La puissance d'un test est liée à la taille de l'échantillon, plus l'échantillon est petit, plus le risque de passer à côté d'une association est grande.

Le risque α est aussi appelé risque de première espèce. Il correspond donc pour un test à un seuil de rejet de H_0 tout en acceptant le risque de la rejeter à tort.

- **Construction d'un test d'hypothèse**

Pour comparer une hypothèse H_0 à l'hypothèse alternative H_1, nous devons établir un critère de décision permettant de choisir entre les deux hypothèses. Pour cela nous procédons par étapes :

(i) On pose l'hypothèse H_0 l'hypothèse nulle, dans notre cas que notre SNP n'est pas associé à notre phénotype, et H_1 l'hypothèse alternative, notre SNP est associé à notre phénotype ;

(ii) On détermine la statistique liée à l'hypothèse H_0 dont on connaît la loi de distribution ;

(iii) On détermine la zone de rejet en fonction de H_1 et du seuil de significativité ;

(iv) On évalue la statistique de l'échantillon à partir de la loi de distribution ;

(v) Conclure sur le test en fonction de la statistique et de la zone de rejet.

- **Exemple d'un test d'association entre un SNP et un phénotype**

Dans cette partie, nous allons voir l'illustration d'un test d'hypothèse dans le cadre d'un exemple. On cherche à savoir si notre SNP A sous le modèle génétique dominant pour l'allèle a_1 est associé à notre phénotype binaire (atteint/non atteint) dans nos populations cas et témoins.

Nous pouvons alors construire un tableau de contingence représentant notre exemple :

	Cas	Témoins
Porteurs a_1	$N_{a_1a_1}^{Cas} + N_{a_1a_2}^{Cas}$	$N_{a_1a_1}^{Témoins} + N_{a_1a_2}^{Témoins}$
Non Porteurs a_1	$N_{a_2a_2}^{Cas}$	$N_{a_2a_2}^{Témoins}$

Tableau 3 : Exemple de tableau de contingence cas/témoin sous un modèle dominant

Étape 1 : On pose l'hypothèse nulle H_0, la répartition de nos porteurs d'allèles a_1 est identique dans nos population cas et témoins. On pose l'hypothèse H_1, la répartition de nos porteurs d'allèle a_1 est différente entre nos populations cas et témoins.

Étape 2 : Nous avons donc une distribution qui suit une loi du $\chi 2$ à un degré de liberté sous H_0.

Étape 3 : On détermine alors la zone de rejet de notre distribution en fonction de H_1 et de notre seuil de significativité.

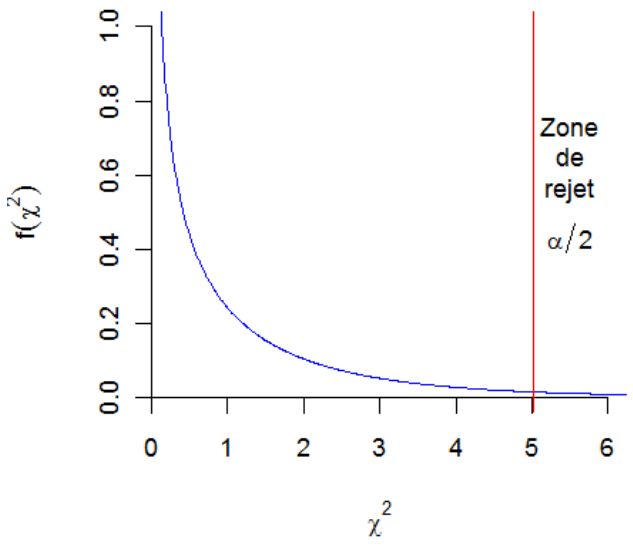

Figure 12 : Fonction de distribution d'une courbe de χ^2 à 1 degré de liberté dans le cadre d'un test bilatéral

Étape 4 : Nous pouvons donc calculer la statistique de notre échantillon avec la formule suivante

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

avec *i* les modalité de la première variable à savoir le fait de porter oui ou non l'allèle a1

avec *j* les modalités de la seconde variable à savoir le fait d'être un cas ou un témoin

avec *O* la valeur observée

avec *E* la valeur attendue obtenue par le produit des sommes marginales divisée par la somme totale par exemple $E_{Porteur\,a_1 / Cas} = \frac{N_{Cas} \times N_{Porteurs\ a_1}}{N_{Cas} + N_{Témoins}}$.

Étape 5 : A partir de seuil de rejet établi à l'aide du risque a choisi, on peut alors conclure pour le test. Deux cas sont possibles.

Soit la valeur du χ^2 calculée est en dessous du seuil de rejet et alors on ne rejette pas H_0. H_0 est alors "acceptée par défaut" sans pour autant qu'il ait été prouvée qu'elle soit vraie,

faute de preuve suffisante pour la rejetée. On conclurait donc que la distribution de notre SNP sous le modèle dominant n'est donc pas associée à notre phénotype dans notre échantillon.

Soit la valeur du χ^2 calculée est au-delà de la zone de rejet et alors on rejette H_0. On rejette donc H_0 au profit de H_1 au risque de commettre une erreur de type I établi. On conclurait donc que la distribution de notre SNP sous le modèle dominant est associée à notre phénotype avec un risque α de se tromper dans notre échantillon.

Ceci n'est qu'un exemple d'un test de comparaison de distribution. Il appartient à l'expérimentateur de vérifier si les conditions de validité du test sont remplies. Ce test s'applique à notre échantillon, de part sa structure (phénotype binaire et répartition de notre SNP par classe). Des situations différentes telles que des phénotypes autres que binaire ou bien d'autres structures de répartition de SNP entrainerons des statistiques différentes. Le cas représenté n'en est qu'une des possibilités les plus simples.

- **P-value**

Après avoir décidé ou non du rejet de H_0, nous pouvons déterminer la p-value de notre observation qui correspond à la probabilité d'observer une conformation au moins aussi extrême que celle observée sous l'hypothèse H_0. La p-value constitue donc une mesure de significativité pour un test : plus la p-value est petite, moins le hasard semble avoir eu d'effet dans notre observation.

La p-value permet donc de trier les différents tests en fonction de leur significativité. La p-value permet de décider du rejet de H_0 pour le test d'un SNP. Elle permet aussi de trier les SNPs testés en fonction de leur éloignement par rapport à la distribution attendue sous H_0.

Néanmoins, la p-value n'informe que sur la distance vis à vis d'une distribution attendue sous H_0, elle n'en précise pas toujours l'effet, à savoir si notre SNP dans son mode étudié possède un effet protecteur ou aggravant vis à vis de notre phénotype (dans le cadre d'un test bilatéral).

- **Odds Ratio (rapport de cotes)**

Bien qu'un simple coup d'œil aux fréquences alléliques ou génotypiques suffit à nous renseigner sur le sens (protecteur/aggravant) d'une association, nous pouvons aussi chercher à ordonner ces associations en fonction de leur intensité. L'Odds Ratio (OR) correspond comme son nom l'indique à un rapport des chances dans notre tableau de contingence 2vs2 (tableau 2). Il n'est pas en lui-même un test statistique mais une unité de mesure de l'intensité

de l'association ou tout comme son nom l'indique encore l'Odds Ratio entre notre SNP et notre trait. En reprenant les notations introduites dans le tableau 2, on a :

$$Odds\ Ratio = \frac{\frac{A}{A+B} \Big/ \frac{B}{A+B}}{\frac{C}{C+D} \Big/ \frac{D}{C+D}} = \frac{AD}{BC}\ défini\ sur\ [0,+\infty[$$

avec un intervalle de confiance à 95% définit par :

$$\left]\exp(\log(OR)) - 1.96 * \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}; \exp(\log(OR)) - 1.96 * \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}\right[.$$

L'Odds Ratio pour un SNP sans effet est égal à 1. Plus l'OR se rapprochera des bornes de définition, plus l'effet sera important, prévenant le trait étudié lorsqu'il tend vers 0 et le favorisant lorsqu'il tend vers $+\infty$. A noter que l'OR n'a de significativité que si son intervalle de confiance ne comprend pas 1, sans quoi il est impossible de conclure.

- **Régression linéaire**

Lorsque nous avons un phénotype quantitatif, nous sortons alors du cadre de l'exemple décrit dans le tableau 2. Nous pouvons toutefois chercher une association entre notre variable à expliquer (notre phénotype quantitatif) et notre variable explicative (nos génotypes ou nos allèles selon le modèle génétique utilisé). L'idée étant de représenter la relation entre nos deux éléments comme une fonction affine :

$$Y \approx f(X)$$

avec Y notre variable à expliquer

et X notre variable explicative.

Nous possédons alors n couples d'observation y_i et x_i disponibles où il faudra ajuster f (notre fonction) parmi F (une classe de fonctions) dans laquelle nous supposerons que se trouve la vraie fonction inconnue. Le problème peut donc se résumer comme un problème de minimisation de coût :

$$\arg\min_{f \in F} \sum_{i=1}^{n} coût(y_i - f(x_i))$$

La classe de fonction F la plus naturelle est la fonction affine de l'ensemble des réels vers l'ensemble des réels appelée régression linéaire simple.

On peut alors représenter l'ensemble des *n* observations par l'équation suivante :

$$\forall i \in \{1,\ldots,n\} \quad y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

avec y_i notre variable à expliquer considérée comme une variable aléatoire

avec x_i notre variable explicative considérée comme une variable non aléatoire

avec ε_i un terme d'erreur supposé centré, de même variance, non corrélées entre elles et aléatoires

avec β_1 et β_2 sont les coefficient de la régression linéaire, ils sont les paramètres inconnus (mais non aléatoires) du modèle que nous allons donc estimer.

On peut alors estimer les coefficients β_1 et β_2 avec la méthode des moindres carrés ordinaire. Il est à noter que l'on peut inclure d'autres variables explicatives (dans notre cas nous les appellerons covariables) dans la fonction de régression, on parle alors de régression linéaire multiple. Chacune des covariables seront alors des variables explicatives (non aléatoire) avec leur coefficient de régression (inconnu mais non aléatoires).

- **β coefficient de régression linéaire**

Le coefficient de régression apparait lors des régressions utilisées selon les données disponibles ainsi que des hypothèses testées. Chacune des variables utilisées pour prédire notre trait possède donc son propre coefficient et nous obtenons aussi un coefficient β_0 qui correspond à un état basal sans implication de nos variables.

Il se définit sur $]-\infty;+\infty[$, il est considéré comme sans effet lorsqu'il est proche de 0, protecteur lorsqu'il est négatif et aggravant lorsqu'il est positif.

1. Facteurs de confusion

Lorsque l'on teste l'association entre un SNP et un trait, d'autres variables (covariables) peuvent avoir une influence sur cette association, ces covariables doivent alors être prises en compte lors des analyses.

- **Covariables**

Plusieurs facteurs extrinsèques peuvent influencer le trait étudié. Ainsi des facteurs environnementaux, comportementaux peuvent avoir un impact sur le phénotype étudié sans avoir de lien avec le SNP. Par exemple dans les études sur le cancer, des facteurs tels que le tabac, l'alimentation, ou l'exposition aux radiations peuvent entrer en jeu et doivent être pris en compte lors des tests d'association entre les SNPs et le cancer.

- **Stratification**

En génétique, la stratification est un nom qui permet de désigner un phénomène de "différenciation" dû à des phénomènes de migrations ancestrales. Par exemple, deux populations avec deux histoires de migrations ancestrales distinctes seront différentes même si elles partagent le même trait (cas/contrôle). Ces différences peuvent être observées à l'échelle continentale mais aussi à celle des pays voisins comme l'Europe[52] (figure 13). Ces différences doivent être prises en compte lors de l'analyse car elles peuvent devenir un facteur de confusion, les associations observées n'étant pas liées au phénotype étudié mais à une structure migratoire différente au sein de notre population.

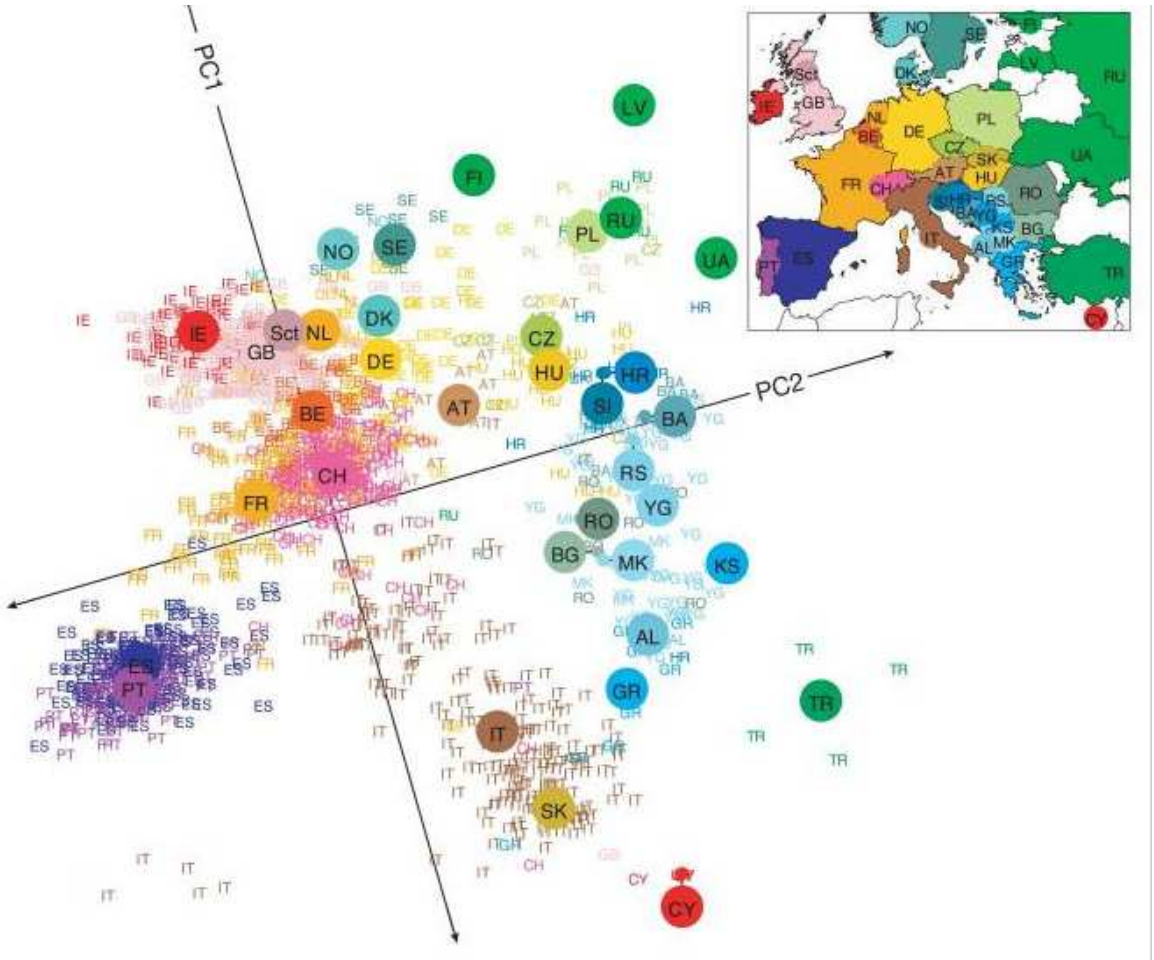

Figure 13: Représentation des deux premiers axes de l'analyse par composantes principales

Source: Novembre et al. 2008 PMC©

En pratique, cette histoire migratoire peut être déterminée par des programmes tels que STRUCTURE [53, 54] ou bien EIGENSTRAT [55]. Par la suite, les sujets considérés comme étant atypiques, c'est-à-dire loin des autres, sont retirés de l'étude et les individus restants sont

alors analysés en prenant en compte les axes principaux comme covariables associées à leur histoire migratoire.

e. Contrôle de qualité de l'analyse : Q-Q plot

Après avoir obtenu une p-value pour chacun des SNPs, il faut s'assurer que lors de l'analyse, des facteurs de confusion inconnus n'aient pas été oubliés. Pour cela, on utilise le Diagramme Quantiles-Quantiles qui revient à comparer la distribution des p-values observées par rapport à la distribution théorique des p-value attendues (sous l'hypothèse H_0, les p-values selon une loi uniforme sur [0,1]). Les p-values observées doivent se superposer à la droite représentant la distribution théorique hormis bien entendu des signaux d'association qui devraient s'éloigner de cette droite. Une dissymétrie peut signifier qu'il y a un facteur de confusion qui n'a pas été pris en compte lors de l'analyse. Néanmoins, il n'est pas toujours possible d'identifier ce facteur de confusion. On peut alors réajuster les p-values en utilisant le facteur d'inflation génomique [56, 57] basée sur la médiane de la statistique du χ^2.

f. Recherches post-association

Même si un SNP a été déclaré associé au trait étudié, il n'est peut-être pas pour autant le SNP causal. Bien que seule l'expérimentation (tests *in vitro*, tests animaux, tests cliniques) puisse répondre à cette question, plusieurs approches peuvent apporter de la lumière sur cette question : la recherche bibliographique, les analyses haplotypiques, les techniques d'imputation ainsi que les éventuelles réplications ou méta-analyses.

1. Bases de données

A la découverte d'un SNP associé significativement avec un phénotype, nous pouvons rechercher plus d'information sur ce SNP. Le premier point est la bibliographie, chaque SNP étant renseigné dans la base de données dbSNP, il peut être décrit par divers éléments tels que sa localisation sur un gène, son impact sur un ARN ou sur une protéine. De plus, il faut prendre en compte le fait que ce SNPs ne soit qu'un TagSNP qui marque éventuellement le SNP causal absent sur la puce de génotypage.

Nous pouvons ensuite déterminer s'il appartient à un gène et plus précisément s'il correspond à une mutation de la protéine ou toute autre modification pouvant influer sur la fonction du gène. Ces informations sont référencées dans la base de données dbSNP

(http://www.ncbi.nlm.nih.gov/projects/SNP/). De plus même si un SNP n'impacte pas directement sur la structure du gène il peut toutefois impacter sur l'expression de gènes de par sa localisation dans des régions promotrices [58, 59]. Enfin, il ne faut pas oublier que le SNP d'intérêt peut aussi être en déséquilibre de liaison avec des SNPs fonctionnels.

2. Analyse des haplotypes

L'analyse des haplotypes est importante car en réalité plusieurs SNPs peuvent impacter le trait, étant transmis sur le même chromosome. Deux allèles de deux SNPs peuvent donc avoir un effet combiné bien plus important que les deux SNPs considérés isolément comme dans le cas du récepteur 2-adrénergique beta (*ADRB2*) [60].

Les haplotypes présentent un problème dans les GWAS : l'inflation combinatoire. En effet, il y a une combinaison quasi infinie de SNPs possibles pour les haplotypes, même si on se limite aux SNPs voisins d'un même chromosome. De nombreuses études génome entier sont consacrées aux haplotypes [61, 62] mais elles sont le plus souvent limitées aux haplotypes de régions contenant des SNPs à effet individuel très fort. Toutefois, l'analyse des haplotypes permet aussi de répondre aux questions plus générales sur l'origine des gènes [63] et de retracer l'histoire migratoire de diverses populations [64].

3. Imputation

Plus récemment, les progrès de la bioinformatique ainsi que l'augmentation des ressources génomiques disponibles (panels de génomes de référence), ont permis d'imputer les génotypes des SNPs absents des puces à partir des données disponibles. Le procédé vise à reconstruire les haplotypes des sujets et à les aligner sur un panel d'haplotypes de référence afin de reconstituer les génotypes manquants.

Plusieurs programmes permettent d'imputer les génotypes tels que MACH [65] et IMPUTE [66, 67]. Par expérience, l'imputation ne permet pas, en général, de trouver de nouvelles régions, mais affine et précise les niveaux d'association dans les régions les plus fortes déjà identifiées. L'imputation sert principalement à pouvoir mélanger des données génotypiques issus de protocoles différents et ne comprenant pas forcément les mêmes jeux de SNPs. L'imputation peut aussi servir à préciser les signaux d'une région avec un SNP d'intérêt en reconstituant les SNPs de la région qui ne sont pas présents dans notre jeu de

donnée en espérant identifier un SNP dont l'impact potentiel sur notre phénotype soit plus flagrant que celui de notre SNP d'intérêt.

4. Réplication & méta-analyse

Pour confirmer un résultat d'association dans une GWAS, on peut tenter de le répliquer dans une autre étude. Cela consiste à rechercher si la p-value de notre SNP d'intérêt se révèle aussi significative dans une étude impliquant sur des traits similaires. Néanmoins, il est souvent très difficile de répliquer une étude de par le fait de trouver une population satisfaisant des critères d'inclusion identiques ou bien tout simplement de retrouver le SNP dans une autre étude déjà réalisée. Par exemple, les jeux de SNPs sont différents entre Illumina et Affymetrix, même si l'imputation peut combler cette lacune en tentant de reconstruire artificiellement le SNP manquant. Enfin si l'étude est novatrice, il sera bien entendu impossible de répliquer le signal tant que d'autres études n'ont pas été menées.

Il est aussi possible d'approfondir la comparaison des études en effectuant une méta-analyse, à savoir calculer une p-value globale sur plusieurs études pour chaque SNP. Après avoir vérifié que nos données sont homogènes et donc comparables entre différentes études, on peut ensuite estimer une p-value globale. Plusieurs méthodes existent permettant soit de calculer la probabilité que ces SNPs apparaissent ainsi dans les différentes études par une Statistique de Fisher [68] soit de prendre en compte le sens de l'association (OR ou Beta) par le biais d'un Z score [69]. Ces méta-analyses permettent parfois d'identifier des signaux intéressants sans qu'ils aient été nécessairement significatifs dans chacune des études individuelles.

5. Redondance de l'information et tests multiples

Dans ce chapitre, seront présentés la problématique du contrôle du risque de première espèce dans le cadre des tests multiples et différentes solutions apportées pour ce problème. Nous introduirons ensuite l'entropie de Shannon ainsi que son application sur la redondance de l'information.

a. Correction des tests multiples

1. Problématique

Nous avons vu précédemment comment construire une étude génome entier ainsi que la procédure pour tester une association entre un SNP et le phénotype étudié (cf. chapitre 4 : Analyse d'association génome entier). Nous avons vu que lors d'un test, nous avons un risque de rejeter à tort H_0, appelé risque α ou bien probabilité d'erreur de 1ère espèce. Lors d'un test unique, ce risque est généralement fixé à 5%.

Cependant, ce risque α n'est valable que pour un seul test. Si l'on répète ce test pour plusieurs SNPs, nous ne pouvons garder ce même seuil pour tous les SNPs car nous aurons alors une inflation du nombre de faux positifs. Par exemple, si l'on teste 100 SNPs indépendants dont nous savons qu'ils n'ont aucune association avec notre phénotype, avec ce même seuil de 5% nous nous attendons à rejeter H_0 à tort pour 5 SNPs.

Le nombre de faux positifs pour un seuil donné augmente bien entendu avec le nombre de tests effectués. Le risque α doit être corrigé afin d'avoir un niveau de risque étendu à l'ensemble des tests acceptable. Plusieurs corrections dites des "tests multiples" ont vu le jour afin de palier ce problème.

2. Méthodes de correction pour la problématique des tests multiples

Plusieurs méthodes ont été développées afin de corriger ce problème de tests multiples. Nous allons voir tout d'abord les méthodes qui permettent de contrôler le risque α

sur l'ensemble de l'expérimentation. De plus, lorsque nous travaillons sur des données génomiques, nous avons l'intuition que le déséquilibre de liaison rend les tests dépendants entre eux. Nous verrons ensuite les différentes méthodes qui permettent de déterminer le nombre de SNPs indépendants pour un jeu de données génomiques.

Sidak - Bonferroni

La correction de Sidak-Bonferroni est la correction la plus utilisée car la plus robuste mais elle est réputée très conservatrice. Pour un test, nous pouvons déclarer que la probabilité de ne pas rejeter à tort H_0 s'écrive :

$$1-\alpha .$$

Pour N évènements indépendants, nous pouvons alors déclarer que la probabilité de ne pas commettre d'erreur de première espèce dans aucun des N tests est :

$$(1-\alpha)^N .$$

La probabilité de faire au moins une erreur première espèce sur ces N tests est donc le risque α pour N tests, ou appelé aussi α pour la famille de test, s'écrit donc :

$$\alpha_{N\,tests} = 1-(1-\alpha_{1\,test})^N .$$

Qui peut aussi se réécrire : $\alpha_{1\,test} = 1-(1-\alpha_{N\,tests})^{1/N}$.

Ce qui correspond à la correction de Sidak et permet donc de réajuster le seuil α pour un test afin de garder un seuil α étendue à toute la famille. Néanmoins cette formule a pour inconvénient de nécessiter le calcul d'une puissance fractionnelle ce qui était moins aisé avant les progrès de l'informatique. C'est pourquoi Bonferroni a approximé cette expression en :

$$\alpha_{N\,tests} \approx \frac{\alpha_{1\,test}}{N}\ soit\ donc\ \alpha_{1\,test} \approx \frac{\alpha_{N\,tests}}{N} .$$

C'est cette dernière expression qui est appelée correction de Bonferroni. Elle est la plus utilisée de par sa simplicité de mise en oeuvre et permet aussi de corriger les p-values. Il est important de noter que pour 2 tests ou plus nous avons :

$$\frac{\alpha_{N\ tests}}{N} < 1-(1-\alpha_{N tests})^{1/N}.$$

Cela signifie que la correction de Bonferroni via son approximation produit un seuil plus stringent que celui de la correction de Sidak. De plus, il est aussi important de noter que ces corrections supposent l'indépendance des tests ce qui n'est en général pas le cas au sein du génome à cause du déséquilibre de liaison existant entre les SNPs.

Néanmoins, l'hypothèse erronée d'indépendance entre les tests peut entrainer à tort le rejet de certains SNPs. C'est pourquoi plusieurs alternatives ont été proposées afin de pallier cette correction sur-conservatrice.

Taux de fausse découverte FDR

La méthode du taux de faux positifs (False Discovery Rate en anglais, FDR) permet d'évaluer parmi les tests considérés comme les meilleurs, le taux de tests que nous déclarons à tort comme étant significatifs [70]. Cette méthode est moins contraignante que la correction de Bonferroni car elle nous permet de contrôler le taux de faux positifs parmi les résultats intéressants mais ne passant pas le critère de Bonferroni. Elle produit alors pour chaque test une valeur appelée q-value qui correspond donc au taux de faux positifs parmi les p-values meilleures ou égales à celle testée, i.e. les SNPs dont la p-value est plus basse ou égale. Le seuil de 25% est couramment utilisé ce qui signifie qu'en dessous de ce seuil, on considère qu'un quart des signaux est déclaré significatif à tort mais aussi que les trois quarts sont déclarés significatifs à raison, ce qui peut fournir des informations potentiellement intéressantes pour les biologistes.

Cette méthode connait de nombreuses déclinaisons [71] en ne se focalisant que sur une fenêtre locale de tests [72, 73]. Néanmoins, ces déclinaisons sont intrinsèquement liées à la distribution des p-values de l'étude considérée. Elles souffrent donc de leur absence de portabilité d'une étude à l'autre ce qui peut être un frein à la réplication des signaux, même si elles peuvent être recalculées dans le cas d'une méta-analyse.

Test de permutations

Cette méthode revient à calculer la probabilité d'observer une conformation similaire ou plus extrême sous l'hypothèse H_0 à savoir qu'il n'y ait pas d'association entre la variable explicative et la variable à expliquer, i.e. la probabilité d'observer le tableau de contingence

entre le SNP et le trait par hasard. Pour cela, la méthode va construire aléatoirement des conformations respectant les sommes marginales. En pratique, cela revient à permuter les statuts, par exemple cas et témoins, au hasard. Pour chaque test aléatoire nous allons donc obtenir une statistique de test. Il suffit alors de compter le nombre de statistiques équivalentes ou plus extrêmes que celle observée et de diviser ce nombre d'observations par le nombre de permutations.

Ceci permet donc d'obtenir une p-value empirique. Concrètement pour un SNP donné, nous allons donc mélanger les sujets des populations cas et contrôle pour retirer au hasard une population cas et une population contrôle, et pour chaque permutation recalculer une statistique de test. Il suffit alors de comptabiliser le nombre de p-values équivalentes ou plus extrême que la p-value de base et de le diviser par le nombre de permutations pour obtenir la p-value empirique du SNP. Le nombre de permutations au moins aussi important que le nombre de variables testées afin de franchir le seuil de Bonferroni. Ces permutations permettent de s'affranchir de la non-indépendance des tests.

Cette méthode nécessite de recalculer un nombre aussi important de p-values que de permutations et ce pour chaque SNP testé afin de parvenir au seuil de Bonferroni. De nombreux logiciels ont implémenté cette technologie avec le calcul des statistiques usuelles [74-76]. Avec les progrès en informatique, la méthodologie est applicable mais nécessite de longs temps de calcul surtout pour des statistiques complexes à calculer. Ces permutations restent cependant très intéressantes pour vérifier un résultat.

Ces différentes méthodes se sont focalisées sur la correction du seuil α pour l'ensemble des tests effectués ou bien sur une meilleure définition des signaux comme le taux de fausses découvertes ou bien le re calcul des p-values empiriques par permutations. Toutefois, ces méthodes ne s'attardent pas sur la non-indépendance des tests comme on peut intuitivement le supposer lorsque l'on traite des données génomiques et a fortiori lorsque des SNPs sont en déséquilibre de liaison. Une seconde catégorie de méthodes ont été développées afin de déterminer le nombre de tests indépendants pour un jeu de données génomiques. L'idée de ces méthodes est donc de déterminer le Meff qui correspond au nombre de tests indépendants effectués et ainsi pouvoir s'en servir pour la correction de Sidack-Bonferroni.

Calcul du Meff

L'idée du Meff est qu'un SNP en déséquilibre de liaison total avec un autre SNP est un poids mort pour l'analyse statistique puisque nous allons tester deux fois les mêmes SNPs (au point de vue variable et modalités). En 2008, Duggal [77] propose une méthode basée sur le déséquilibre de liaison, plus précisément la mesure du D'. En se reposant sur elle, ils définissent des blocs de SNPs corrélés où chacun des SNPs est reliés aux autres avec un seuil de D' > 0,70. Le Meff s'obtient alors par sommation du nombre de blocs obtenus. Toutefois cette méthode est dépendante du seuil de D' utilisé faisant presque passer du simple au double le Meff obtenu.

D'autres méthodes sont basées sur des matrices de déséquilibre de liaison entre toutes les paires de SNPs possibles ou bien de corrélation sur les comptes d'allèles entre les paires de SNPs, ce qui est une méthode approximative et rapide de calculer le déséquilibre de liaison. Ces méthodes suivent toutes le même modèle :

Étape 1 : calcul d'une matrice de corrélation

Dans cette étape, on calcule donc le déséquilibre de liaison pour chaque paire de SNPs. Selon les études et les besoins les mesures de corrélations/déséquilibre de liaison peuvent être différentes. La méthode la plus courante est celle du déséquilibre de liaison composite [78] qui correspond à la prise en compte des SNPs en fonction de leur compte d'allèles et de calculer le coefficient de corrélation de Pearson entre les deux SNPs.

Étape 2 : calcul du Meff

A partir de la matrice des corrélations, on peut extraire les valeurs et vecteurs propres de la matrice. Ces valeurs propres sont utilisées pour calculer le Meff suivant différentes formules. Les valeurs propres peuvent être utilisées en tant que telles ou bien pour leur variance dans ce cas, on suppose que les valeurs propres ont pour moyenne 1.

Étape 3 : calcul du nouveau seuil

Enfin, après avoir défini le Meff, le nouveau seuil est calculé soit par la correction de Sidack soit par celle de Bonferroni

Le tableau ci-dessous résume les différences entre les différentes méthodes :

Méthode	Corrélation pour une paire de SNPs	Calcul du Meff
Gao et al.[79-81]	Composite LD Corrélation de Pearson	seuil C fixé à 95% $\sum_{i=1}^{x} \lambda_i / \sum_{i=1}^{M} \lambda_i > C$ $Meff = x$
Cheverud et al.[82]	Composite LD Corrélation de Pearson	$Meff = M(1 - \frac{(M-1)V_{\lambda obs}}{M^2})$ $avec\ V_{\lambda obs} = \sum_{i=1}^{M} \frac{(\lambda_i - 1)^2}{(M-1)}$
Li et al.[83]	Composite LD Corrélation de Pearson	$Meff = \sum_{i=1}^{M} f(\lvert\lambda_i\rvert)$ $avec\ f(x) = I(x \geq 1) + (x - \lfloor x \rfloor), x \geq 0$ $avec\ I(x) = 1\ pour\ x \geq 1\ et\ I(x) = 0\ pour\ x < 1$
Nyolt et al.[83]	$\Delta = \frac{f_{a1b1}f_{a2b2} - f_{a1b2}f_{a2b1}}{\sqrt{f_{a1}f_{a2}f_{b1}f_{b2}}}$	$Meff = 1 + (M-1)(1 - \frac{V_{\lambda obs}}{M})$ $avec\ V_{\lambda obs} = \sum_{i=1}^{M} \frac{(\lambda_i - 1)^2}{(M-1)}$
Galwey et al.[84]	Composite LD Corrélation de Pearson	$Meff = \frac{(\sum_{i=1}^{M} \sqrt{\lambda_i})^2}{\sum_{i=1}^{M} \lambda_i}$

Tableau 4 : Récapitulatif des différentes méthodes de détermination des Meff

M correspond au nombre de SNPs total

λ correspond à une valeur propre obtenue par décomposition spectrale

D'autres méthodes existent pour corriger le problème des tests multiples. En dehors du calcul du nombre de tests indépendants, nous pouvons définir des seuils universels selon la technologie utilisée. En procédant à des simulations de données et des permutations, plusieurs études [85, 86] ont réussi à déterminer de nouveaux seuils de significativité selon la puce utilisée. Toutefois ces seuils ne sont valables que si la puce ou bien la technologie utilisée est présente dans leurs études et aussi que la population testée est similaire à celle utilisée ou

simulée dans leurs études. Ces seuils sont donc de bonnes indications toutefois elles deviennent très vite obsolètes au vu des périodes de commercialisation des puces de génotypage.

Par la suite, une méthode pour estimer la liaison de dépendance globale entre différents SNPs sera détaillée. Auparavant, différents concepts issus de la théorie de l'information seront présentés.

b. Entropie

1. Définition

L'entropie de Shanon [87] est un concept introduit pour la théorie de l'information en 1948 pour caractériser l'information entre un transmetteur et un receveur d'informations au début de l'informatique. Par exemple, si le message est redondant l'information apportée par chacune des répétitions est nulle. L'entropie permet alors de quantifier le désordre, ou incertitude, associée à une variable dont l'unité de mesure est le bit.

Pour une variable X = $\{x_1, x_2, ..., x_n\}$ de probabilité d'observation $p_1, p_2, ..., p_n$ l'entropie se calcule par la formule suivante :

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i .$$

L'entropie d'une variable est une fonction dépendant de ses modalités et de leurs probabilités d'observation. Dans le cas d'un SNP, en faisant abstraction des bases nucléotidiques, nous pouvons résumer les modalités d'un SNP *A* à ses deux allèles a_1 et a_2, dont les probabilités d'observations sont les fréquences alléliques f_{a_1} et f_{a_2}. L'entropie du SNP *A* se calcule grâce à la fonction suivante :

$$H(A) = -fa_1 \log_2(fa_1) - fa_2 \log_2(fa_2) .$$

A noter que l'entropie peut être déclinée pour tout polymorphisme, tant que nous connaissons ses fréquences alléliques.

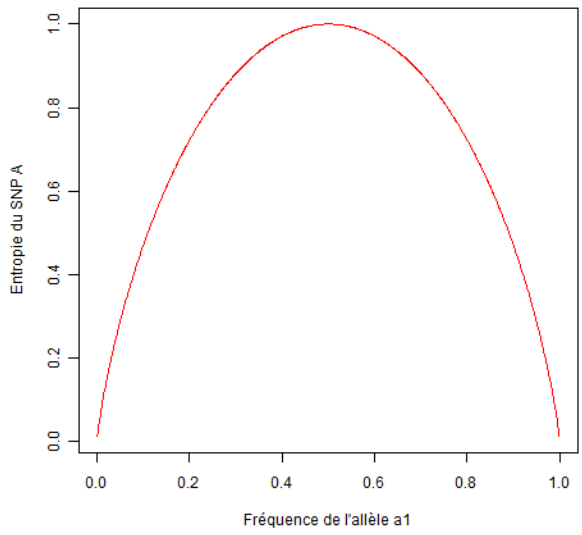

Figure 14 : Courbe d'entropie d'un SNP bi-allélique en fonction de la fréquence allélique

L'entropie possède quelques propriétés intéressantes dont je vais présenter une liste ci-dessous :

- L'entropie est **non négative,** $H(A) \geq 0$ est nulle si et seulement si la probabilité d'une des modalités est égale à 1 et donc que toutes les autres modalités ont une probabilité nulle, soit donc que notre SNP *A* est un SNP monomorphique et qu'un seul allèle est présent ;

- L'entropie est **symétrique,** ceci revient à dire que l'on peut permuter les modalités sans changer l'entropie de la variable tant que chaque modalité garde sa probabilité d'observation (figure 14) ;

- L'entropie est **maximale**, lorsque toutes les modalités de la variables sont équiprobables ;

- Cas particulier d'une variable binaire, la probabilité d'observation d'une des modalités conditionne la probabilité d'observation de l'autre modalité $f_{a_2} = 1 - f_{a_1}$ on peut donc écrire que $H(f_{a_1}, f_{a_2}) = H(f_{a_2}, f_{a_1}) = H(f_{a_1}, 1 - f_{a_1})$. Cette symétrie dans la distribution des probabilités d'observation explique la symétrie de la courbe (figure 14). Du fait que la

variable n'a que deux modalités observables, l'entropie est maximale lorsque les modalités sont équiprobables, soit lorsque $f_{a_1} = f_{a_2} = \frac{1}{2} = 0.5$, son entropie est alors égale à 1.

2. Déclinaison

On peut aussi utiliser l'entropie de Shannon sur plusieurs variables en connaissant leurs modalités et leurs probabilités d'observations respectives. Dans le cadre d'un couple de variables, les modalités du couple correspondent au produit cartésien des modalités des variables. Dans le cas de deux SNPs *A* et *B*, le couple formé (A, B) a alors pour modalités a_1b_1, a_1b_2, a_2b_1 et a_2b_2 et leur probabilité d'observation respective $fa_1b_1, fa_1b_2, fa_2b_1$ et fa_2b_2. L'entropie du couple (A, B) peut donc se calculer suivant la même formule énoncée pour une simple variable.

Avant d'aborder l'entropie conditionnelle, je vais rappeler quelques éléments de probabilités conditionnelles :

On a $X=\{x_1, x_2, ..., x_n\}$ et $Y=\{y_1, y_2, ..., y_m\}$ deux variables discrètes aléatoires avec des probabilités d'observations jointes et individuelles on peut donc écrire pour le couple de variables :

$$p(x_i, y_j) = P\{X = x_i; Y = y_j\}$$

$$p(x_i, y_j) \geq 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) = 1$$

Et pour chaque variable :

$$p(x_i) = P\{X = x_i\} \qquad p(x_i) \geq 0 \qquad \sum_{i=1}^{n} p(x_i) = 1$$

et

$$p(y_j) = P\{Y = y_j\} \qquad p(y_j) \geq 0 \qquad \sum_{j=1}^{m} p(y_j) = 1$$

Nous pouvons alors écrire les probabilités conditionnelles suivantes $p(y_j / x_i) = P\{Y = y_j / X = x_i\}$, $p(y_j / x_i) \geq 0$ pour chaque i, avec $\sum_{j=1}^{m} p(y_j / x_i) = 1, \forall i = 1, 2, ..., n$.

Et inversement pour $p(x_i/y_j)$, nous pouvons alors écrire :

$$p(x_i, y_j) = p(x_i)\, p(y_j / x_i) = p(y_j)\, p(x_i / y_j);$$

$$p(x_i) = \sum_{j=1}^{m} p(x_i, y_j);$$

$$p(y_j) = \sum_{i=1}^{n} p(x_i, y_j).$$

En découle alors l'entropie conditionnelle de la variable Y pour une modalité de la variable X: $H(Y / X = x_i) = -\sum_{j=1}^{m} p(y_i / x_i) \log_2 p(y_i / x_i)$, ainsi que l'entropie conditionnelle de la variable Y sachant la variable X comme $H(Y / X) = \sum_{i=1}^{n} p(x_i) H(Y / X = x_i)$.

L'entropie d'un couple de variables possède des propriétés liées auxvariables qui le composent, pour la suite pour un souci de lisibilité et bien que les propriétés soient aussi applicables pour X et Y, nous nous recentrerons vers notre couple de SNPs (A, B) :

- $H(A, B) = H(A) + H(B / A) = H(B) + H(A / B)$;

- $H(A / B) \leq H(A)$, et est égale lorsque A et B sont indépendants à savoir que $p(a_i, b_j) = p(a_i)\, p(b_j), \forall i, j$;

- $H(A, B) \leq H(A) + H(B)$ et est égale lorsque A et B sont indépendants à savoir $p(a_i, b_j) = p(a_i)\, p(b_j), \forall i, j$.

Nous pouvons alors nous focaliser sur un des cas particuliers des couples de SNPs, à savoir l'haplotype AB.

La théorie de l'information est principalement utilisée dans le traitement de signal mais a connu quelques déclinaisons récentes en génomique notamment dans l'analyse des données [88] afin de sélectionner des SNPs non redondants, mais aussi dans la comparaison de

distribution entre cas et témoins [89] ou bien pour tester les interactions entre gènes [90]. Elle est principalement utilisée pour comparer des distributions ou bien comme critère pour estimer les conformations les plus vraisemblables.

3. Information mutuelle

L'information mutuelle correspond à l'entropie présente dans une variable à laquelle on enlève l'entropie conditionnelle à une seconde variable. Dans le cas de notre couple de SNPs, elle correspond à la part d'information contenue dans un SNP déjà présente dans le second SNP. Elle correspond au déséquilibre de liaison évoqué précédemment et se calcule avec la formule suivante :

$$\begin{aligned} I(A \wedge B) &= H(A) - H(A/B) \\ &= H(B) - H(B/A) \\ &= H(A) + H(B) - H(A,B). \end{aligned}$$

Cette information mutuelle est nulle lorsque nos deux variables sont indépendantes. Cette mesure est équivalente au *D* et permet de mesurer un écart entre une distribution attendue sous hypothèse d'indépendance et une distribution observée. Elle mesure donc le déséquilibre de liaison entre nos deux SNPs. Néanmoins, comme le *D* cette mesure est sensible aux fréquences alléliques.

Plusieurs normalisations ont été proposées pour s'affranchir de cette sensibilité aux fréquences alléliques. En 2002, Nothnagel [91] en introduisait une première avec :

$$\varepsilon(A;B) = \frac{I(A \wedge B)}{H(A) + H(B)} \textit{ définie sur } [0;1[.$$

C'est cette mesure que j'utilise par la suite dans mes travaux.

En 2009, Zhang[92] introduit une autre normalisation de l'information mutuelle avec l'information minimale entre les deux SNPs. Cette normalisation rappelle le *D*' et permet donc de mesurer une relation de redondance imparfaite à savoir qu'un SNP permet de déterminer le second mais que le second ne permet pas de déterminer le premier (cas de seulement 3 haplotypes observables) :

$$MIR(A;B) = \frac{I(A \wedge B)}{\min(H(A);H(B))} \textit{ définie sur } [0;1].$$

Il nous faut aussi introduire un autre concept qui est celui de la redondance. Elle correspond à la fraction d'information redondante et donc superflue par rapport à l'information de nos variables considérées de manière indépendante :

$$R = \frac{I(A \wedge B)}{H(A) + H(B)}.$$

C'est cette redondance qui correspond donc à l'ε de Nothnagel définie précédemment que nous utiliserons comme mesure de déséquilibre de liaison car elle nous permettra de déterminer la fraction indépendante de nos SNPs. Elle nous permettra alors de déterminer la fraction indépendante de la région génomique afin de mieux évaluer le seuil de Bonferroni.

6. Objectifs de ma thèse

Je suis arrivé dans l'équipe de recherche du Pr. Zagury comme stagiaire en 2007 pour développer un outil de visualisation des données génomiques et participer à l'analyse des études génomes entiers sur le projet de Génomique de la Résistance à l'Infection par le VIH-1 (GRIV). A l'heure où les premières analyses génome entier étaient réalisées, j'ai décidé de m'engager sur une thèse portant sur l'exploitation des données génomiques dite de "haut débit", afin d'acquérir et d'appliquer cette méthodologie.

J'ai pu bénéficier de l'expertise forte en haplotypage de l'équipe, ce qui m'a conduit à développer un logiciel mesurant la quantité réelle d'information d'un jeu de données génomiques pour la correction des tests multiples.

Au cours du temps, mon projet de thèse s'est donc profilé sur 2 axes :

1. Développer une méthodologie afin de déterminer la fraction d'information réellement indépendante d'un jeu de données génomique (un ensemble de SNPs dans une population donnée) dans le but d'obtenir une meilleure estimation des seuils de correction statistique ;

2. Exploiter le savoir-faire acquis dans le cadre de l'analyse étude génome entier portant sur l'identification de gènes ayant un impact sur photo-vieillissement.

Matériel & méthodes

1. Données utilisées dans le cadre du développement du logiciel Genetropy

a. Cohortes utilisées

Les cohortes témoins constituent les données réelles que j'ai utilisées pour tester la méthodologie développée. Les données génomiques concernant les cohortes GRIV ainsi que celles de la cohorte DESIR ont été filtrées selon les mêmes critères de qualité que ceux utilisés pour la cohorte SU.VI.MAX. Je n'ai procédé à aucun contrôle de qualité pour les données issues du projet 1000 génomes car nous avons considéré qu'ils avaient déjà été réalisés.

1. Cohorte GRIV

La cohorte GRIV (Génomique de la Résistance face à l'Infection du VIH) est composée de patients présentant un profil extrême dans la résistance au VIH-1 et est composée de 86 sujets Progresseurs Rapides (PR) et 300 Non-Progresseurs à Long Terme (NPLT). Les PR sont définis par une chute de cellules T CD4+ à moins de 300 cellules/mm^3 moins de 3 ans après le dernier test séronégatif. Les NPLT sont des individus séropositifs et asymptotiques depuis plus de 8 ans, et présentant un taux de cellules T CD4+ supérieur à 500 cellules/mm^3.

Le génotypage de la cohorte GRIV a été effectué avec des puces Illumina HumanHap 300 permettant de génotyper les individus sur 317 000 SNPs. Elles ont été élaborées d'après la phase I du projet HapMap. Il avait été estimé que l'ensemble des SNPs fréquents dans la Phase I du projet HapMap pouvait être représenté par environ 240 000 TagSNPs avec un seuil de $R^2 > 0,8$, de plus elles ont été enrichies par environ 8 000 SNPs exoniques.

2. Cohorte DESIR

L'étude DESIR (Data from Epidemiological Study on Insulin Resistance syndrom) consiste en un suivi de 9 ans du développement du syndrome d'insulinorésistance. Le groupe

contrôle utilisé lors des études génome entier avec la cohorte GRIV est composé de 697 participants à ce programme, tous non-obèses, normo-glycémiques, d'origine européenne vivant en France et séronégatifs pour le VIH-1. Elle regroupe 281 hommes et 416 femmes âgés entre 30 et 40 ans. Le génotypage est effectué suivant le même protocole utilisé dans la cohorte GRIV.

3. Projet 1000 génomes

Le projet 1000 génomes est le premier projet avec la volonté de séquencer l'intégralité du génome sur un grand nombre de sujets afin de fournir une source plus détaillée de renseignement sur le génome humain. Il a été rendu possible par les progrès dans les technologies de séquençage qui ont permis de réduire leurs coûts. Le séquençage d'un individu se fait par fragmentation et séquençage de ces mêmes fragments. Pour l'heure, la phase 1 comprend 1 092 individus séquencés avec une basse couverture (2-4X) sur le génome entier ainsi qu'une couverture accrue (50X) sur les régions contenant un gène et comprend 36,6 millions de SNPs, 3,8 millions d'insertions/délétions ainsi que 14 000 délétions larges sur 14 populations dont 379 individus d'ascendance européenne. L'objectif du projet est de séquencer 2 500 individus sur 28 populations différentes.

Dans le cadre du projet Genetropy, j'ai utilisé la version juin 2011 du projet 1000 génomes. Nous avons utilisé leurs haplotypes afin de valider l'algorithme d'EM utilisé pour la phase d'haplotypage du logiciel.

b. Algorithme de Kruskal

L'algorithme de Kruskal nous permet de rechercher un arbre recouvrant de poids minimum dans un graphe connexe valué et non orienté. Dans notre cas, les feuilles correspondent à des SNPs et les arcs au déséquilibre de liaison entre deux SNPs. L'idée est donc de rechercher un arbre de SNPs qui maximise la redondance et donc qui minimise l'information nécessaire à la représentation de l'arbre.

Algorithme :

G: graphe

v: une feuille dans notre cas un SNP

E: l'arbre

E=Ø

pour chaque sommet *v* de *G*

faire créer ensemble *(v)*

trier les arêtes de *G* par ordre décroissant de poids *w*

pour chaque arête *(u, v)* de G prise par ordre décroissant de poids *w* et **tant que** *E* ne recouvre pas *G*

faire si ensemble représentatif *(u)* ≠ ensemble représentatif *(v)*

alors **ajouter** l'arête *(u, v)* à l'ensemble *(E)*

union *(u, v)*

renvoyer *E*

A noter qu'à l'origine, l'algorithme a été crée pour trouver l'arbre recouvrant minimal, mais nous l'avons modifié afin qu'il puisse trouver l'arbre maximal recouvrant étant donné que avons pour but de maximiser la redondance au sein de notre graphe. Cet algorithme nous a donc permis de reconstruire un arbre maximal recouvrant permettant ainsi de maximiser le déséquilibre de liaison entre nos paires de SNPs.

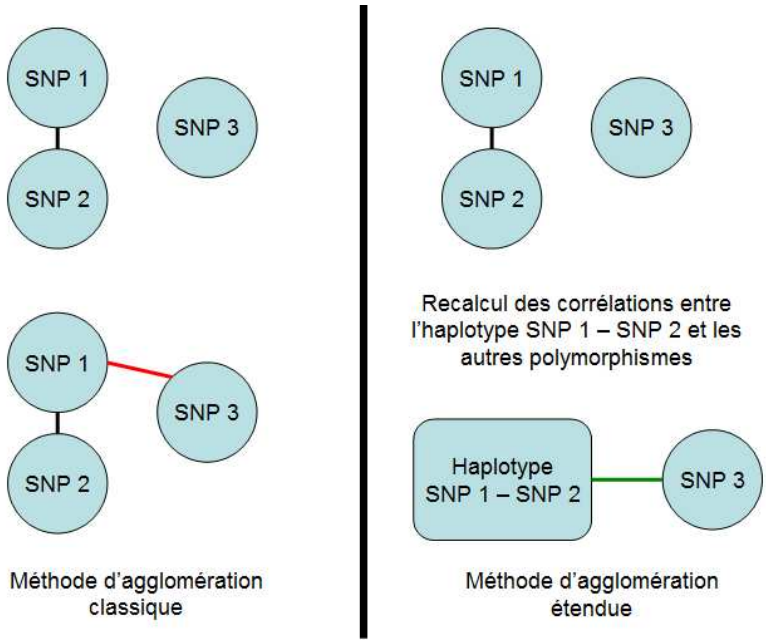

Figure 15 : Méthodes d'agglomération des SNPs

Dans la méthode d'agglomération classique, l'étape de fusion des SNPs n'entraine pas un recalcul des pondérations d'arêtes entre les SNPs. Dans la méthode d'agglomération étendue, l'étape de fusion des SNPs 1 et 2 entraine un calcul des pondérations des arêtes entre l'haplotype SNP 1 - SNP 2 et les autres polymorphismes. Ceci explique pourquoi il y a deux couleurs différentes lors des secondes étapes.

Nous pouvons distinguer deux méthodes de reconstitution de l'arbre. La première est celle que nous appelons la méthode d'agglomération classique où les pondérations des arêtes ne changent pas. Nous considérons alors que les distances entre nos blocs de SNPs sont équivalentes à celles des SNPs. La seconde méthode est celle que nous appelons méthode d'agglomération étendue où les pondérations des arêtes sont recalculées après chaque fusion de SNPs. Cette seconde méthode n'est possible que sur des données phasées intégralement.

c. Calcul des mesures utilisées

Dans notre méthode, nous utilisons plusieurs mesures et introduisons notamment le gain d'information.

Le gain d'information correspond à la compression obtenue entre l'entropie d'un bloc de SNPs pris individuellement par rapport à l'entropie de ce même bloc de SNPs auquel on a ôté sa redondance (de déséquilibre de liaison). L'entropie d'un bloc de SNPs est définie par la

somme des entropies des SNPs qui le compose. La redondance est l'information mutuelle des SNPs en présence. C'est pour cela que l'algorithme de Kruskal à été modifié pour maximiser et non pas minimiser comme il a été initialement conçu. En soustrayant la partie redondante du bloc de SNPs, H(MST), il nous reste alors l'information indépendante contenue dans ce bloc, noté *Ind*. Le nombre de SNPs estimés indépendants (Meff appelé EIS dans la publication), est alors obtenu par une simple règle de 3. Pour des contraintes de charges informatiques, nous avons procédé à un découpage par blocs de SNPs consécutifs pour le jeu de données testé. Nous procédons à deux "passes" décalées puis nous procédons à la moyenne des résultats afin de minimiser les effets de bords.

$$H(SNPs) = \sum_{i=1}^{M} H(SNP_i)$$

$$Ind(SNPs) = H(SNPs) - H(MST)$$

$$Gain = (1 - \frac{Ind(SNPs)}{H(SNPs)}) \times 100$$

$$Meff = M * \frac{Ind(SNPs)}{H(SNPs)}$$

Le gain est une mesure développée pour comparer les performances du logiciel sous différentes conformations telles que des densités de SNPs, effectifs de population, ethnicité de population ou la taille des blocs de SNPs.

2. GWAS sur le photo-vieillissement

a. Vieillissement de la peau

La méthodologie d'analyse génome entier a été utilisée pour identifier des gènes pouvant influencer le photo-vieillissement de la peau. Il est donc important de présenter le contexte de cette étude.

Le vieillissement de la peau est influencé par différents facteurs dont l'âge, l'exposition au soleil et le statut hormonal [93]. Mis à part l'aspect esthétique évident et les conséquences sociales et psychologiques qui lui sont associées, les effets du vieillissement cutané peuvent aussi se traduire par une fréquence augmentée de pathologies cliniques telles que des cancers cutanés. Phénotype visible, le vieillissement de la peau est donc mesurable sans intervention invasive.

Le photo-vieillissement correspond donc à un vieillissement cutané accru lié à des facteurs extrinsèques, par rapport au vieillissement cutané "naturel" uniquement lié quant à lui à des facteurs intrinsèque : âge, phototype (sensibilité naturelle de la peau au soleil)… Parmi les facteurs extrinsèques, on peut noter l'exposition aux ultraviolets [94] (aussi bien au niveau du bronzage solaire que dans les instituts de bronzage artificiel) et la consommation de tabac [95].

Le but de cette étude est donc de caractériser des gènes ayant un impact sur le photo-vieillissement. Le photo-vieillissement est mesurable cliniquement à partir d'une appréciation de l'aspect de la peau en comparaison à une échelle photographique établie en 1994 [96] (figure 15). Dans cette échelle, chaque grade est représenté par trois photographies de référence afin d'illustrer la diversité et la variété des troubles pigmentaires, des rides et du relâchement.

Figure 16 : Illustration photographique de l'échelle du photo-vieillissement

Source : D'après Larnier et al. 1994

b. La cohorte SU.VI.MAX

L'étude SU.VI.MAX [97, 98] (SUpplémentation en VItamines et en Minéraux AntioXydants) est une étude longitudinale conduite en France sur une population adulte d'âge moyen. L'étude SU.VI.MAX avait été initialement développée pour évaluer l'effet d'une supplémentation nutritionnelle quotidienne sur la réduction des problèmes de santé publique, tels que les cancers ou maladies cardio-vasculaires dans les pays industrialisés. La cohorte inclut 13017 volontaires d'âge moyen avec un spectre représentatif de situations sociodémographiques [98]. Le protocole de l'étude SU.VI.MAX a été approuvé par le Comité d'éthique de l'Hôpital Paris-Cochin (CCPPRB n° 706) ainsi que par le "Comité National Informatique et Liberté" (CNIL n°334641). L'étude a été menée conformément aux principes de la Déclaration d'Helsinki.

c. Description de la cohorte des femmes étudiées

L'étude a été conduite sur la période automne - hiver 2002/2003 sur des femmes appartenant à l'étude SU.VI.MAX vivant en région parisienne. Parmi elles (n=2 257), 570 femmes âgées entre 44 et 70 ans ont accepté de participer à cette étude et ont fourni leur consentement éclairé. Les critères d'inclusion étaient les suivants : absence de pathologies connues dermatologiques et absence d'antécédents de procédures esthétiques anti-âge au

niveau du visage. Ces femmes ont été invitées à suivre des consignes spécifiques de soins cutanés notamment l'interdiction d'application de produit de nettoyage ou cosmétiques pour le visage sur les 12 heures avant la visite pour l'étude.

Pour le jour de la visite, elles ont rempli un questionnaire concernant leurs habitudes d'exposition au soleil. Trois photographies haute résolution (2008 x 3032 px) standardisées ont été prises pour chaque participante (une vue frontale de leur face et chacun de leurs profils) avec un appareil photo numérique Kodak DSC 760 avec un objectif 105 mm. L'appareil avait été monté sur un monopode avec une chaise spécialement conçue pour permettre une normalisation de la position du sujet. Les conditions d'éclairage étaient aussi normalisées au moyen de deux lampes symétriques fournissant un spectre de lumière diurne continue, placées à 45° de chaque coté du visage.

Chaque série de photographie a ensuite été examinée par un dermatologue qui a évalué la sévérité des différents signes de vieillissement au niveau du visage, dont le photo-vieillissement global développée par C. Larnier [96]. Dans cette échelle, chaque grade est représenté par trois photographies de référence afin d'illustrer la diversité et la variété des troubles pigmentaires, des rides et du relâchement.

Sur les 570 femmes qui ont participé à l'étude, 68 ont été exclues de l'analyse :

- 18 ont eu une intervention invasive anti-vieillissement ;
- 10 ont été écartées pour être d'ascendance non caucasienne ;
- 1 a été exclue pour cause de quantité insuffisante d'ADN dans le prélèvement ;
- 12 ont été exclues parce que leur ADN a été endommagé ;
- 9 ont été exclues après contrôle qualité du génotypage ;
- enfin 18 ont été exclues pour avoir été déclarées atypiques pour la stratification.

Nous restant alors un total de 502 femmes pour l'analyse génomique dont nous possédons les renseignements pour différentes covariables impactant notre phénotype.

d. Covariables

Nous possédons pour nos 502 femmes, des informations dont l'impact a déjà été démontré sur notre trait :

- l'âge renseigné en années ;
- l'indice de masse corporelle en 3 classes (insuffisance et normalité - IMC < 25kg/m², surpoids - 25 ≤ IMC < 30kg/m² et obésité IMC ≥ 30 kg/m²) [95] ;
- l'exposition au soleil en variable défini comme un score issu de d'une combinaison linéaire de 5 variables (exposition volontaire au soleil, exposition du corps et ou du visage, exposition durant les heures les plus chaudes de la journée, auto-évaluation de l'intensité de l'exposition au soleil tout au long de la vie et importance accordée aux bains de soleil) [95] ;
- la survenue ou non de la ménopause et le cas échéant la prise ou non d'un traitement hormonal de substitution ;
- et le statut tabagique défini en 3 classes (jamais, ancien fumeur, fumeur actuel).

e. Génotypage

Sur les 570 femmes ayant accepté de participer dans l'étude génétique, 529 ont été génotypées avec la puce Illumina Infinium HumanOmni1-Quad contenant 1 140 419 marqueurs. L'ADN génomique (250 ng) a été amplifié, fragmenté, dénaturé et hybridé sur une puce HumanOmni1-Quad pendant au moins 16 heures à 48°C. Les fragments hybridés de manière non spécifique ont été éliminés après lavage et les 795 063 SNPs restants ont été labellisés par fluorescence par extension d'une simple base et ensuite lus avec un scanner IScan (Illumina). Les intensités de fluorescence ont ensuite été normalisées et l'inférence des SNPs a été effectuée à l'aide du logiciel GenomeStudio (v 1.6.3;Illumina). Pour la suite de cette analyse nous nous sommes focalisés uniquement sur les SNPs, les CNVs (n=91 706) ont donc été retirés. De plus, les 2 182 SNPs du chromosome Y ont aussi été retirés puisque la population étudiée est uniquement composée de femmes.

f. Contrôle de qualité du génotypage

Lors du contrôle qualité de l'inférence des SNPs, 9 échantillons ont été éliminés par GenomeStudio (v 1.6.3;Illumina) à cause de leur trop faible taux d'inférence par individu (call rate < 95%). Après quoi les SNPs avec un taux d'inférence par *locus* (call frequency <99%) trop faible ont été ré inférés. Finalement, les individus dont le taux d'inférence par individu était inférieur à 98 % ont été éliminés de l'étude. Ceci correspond au protocole recommandé par Illumina (http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf) afin de minimiser les erreurs d'inférence permettant ainsi d'inférer manuellement les génotypes quand cela était nécessaire.

Toutefois par mesure de précaution, nous avons procédé à un contrôle de qualité supplémentaire afin de palier aux éventuelles défaillances de GenomeStudio. Pour l'analyse, nous avons appliqué les seuils standards de contrôle qualité à savoir :

- données manquantes par individu: inférieur à 2% ;
- données manquantes par marqueur: inférieur à 2% ;
- fréquence allélique mineure par marqueur : inférieur à 1% ;
- seuil au test d'équilibre d'Hardy-Weinberg par marqueur: pvalue < $5 * 10^{-3}$.

Le contrôle qualité se fait lors de l'analyse statistique à l'aide du logiciel PLINK [99].

g. Stratification

Afin de corriger une éventuelle stratification au sein de la cohorte, les génotypes ont été analysés en utilisant le logiciel EIGENSTRAT issu de la suite EIGENSOFT [55], utilisant la méthode d'analyse par composantes principales. Les deux premières passes d'EIGENSTRAT nous ont permis de mettre en évidence 18 individus atypiques qui ont été retirés pour la suite de l'analyse. Une troisième passe nous a alors permis de déterminer les vecteurs propres des individus qui seront par la suite utilisés en tant que covariables, seuls les deux premiers vecteurs propres ont été utilisés afin de ne pas avoir de stratification comme facteur de confusion lors des analyses.

Nous avons aussi utilisé le logiciel STRUCTURE [53, 54] afin de déterminer des individus atypiques néanmoins, nous avons identifié 4 individus atypiques, qui étaient déjà inclus dans ceux identifiés dans les 18 individus atypiques retenus avec EIGENSTRAT.

h. Autres phénotypes

En plus de la détermination du grade de photo-vieillissement établie grâce à l'échelle de Larnier[96], les photos prises lors de la collecte de données ont aussi permis le calcul de scores d'autres phénotypes plus ciblés. Nous avons donc pour chacun de nos sujets, des scores de rides, de lentigines ainsi que de relâchement de la peau au niveau des visages.

Les lentigines correspondent à une hyperpigmentation de la peau avec un contour très bien défini avec une taille variant que quelques millimètres à quelques centimètres de diamètres de couleur allant de marron clair à marron foncé [100]. Les lentigines apparaissent le plus souvent après 50 ans et une exposition chronique au soleil [101].

La sévérité des rides, du relâchement ainsi que des lentigines ont été quantifiée par 3 scores en utilisant les méthodes d'analyse en composantes principales et de régression linéaire[102]. Les valeurs de chaque individu pour chacun des scores ont été transformées pour être comprise entre 0 et 10.

i. Logiciels de traitement des données

L'extraction et le reconditionnement des fichiers de données sont des étapes fréquentes afin de répondre à des questions et besoins ponctuels mais surtout afin d'adapter les fichiers aux différents standards de format des différents logiciels utilisés. Aucun consensus ne s'est imposé en termes de langages de programmation, ni de protocoles. Leurs choix dépendent donc des ressources informatiques disponibles ainsi que des besoins. Mon choix s'est principalement porté sur deux langages de programmation :

- **perl**, langage interprété relativement simple d'utilisation de par son absence de gestion de la mémoire machine, il est répandu dans la communauté biologique et propose de nombreux outils pour la manipulation de texte. Il est portable entre différents systèmes d'exploitation et permet donc une transmission simple de ses scripts. Néanmoins il est bien plus lent qu'un langage interprété et son absence de

gestion de la mémoire machine peut poser problème pour de gros volumes de données ,

- **C++**, langage compilé plus austère que "perl" mais qui autorise la gestion de la mémoire machine, le rendant plus performant en termes de temps de calcul. Il est aussi adaptable au niveau de la mémoire machine permettant la gestion de gros volumes de données. Il est en revanche dépendant du système d'exploitation ainsi que de l'architecture du microprocesseur ce qui freine sa diffusion et transmission.

A titre d'information et d'expérience personnelle, pour un même "besoin simple", le script perl sera plus rapide à mettre en place par rapport à une application en C++, mais le temps d'exécution est 10 fois plus long. Mais la différence entre 10 secondes et 1 seconde reste encore acceptable lors d'une utilisation ponctuelle.

j. Logiciels d'analyse des données

Les données de génotypage ont été traitées et analysées avec le logiciel PLINK [99], boite à outils de la génomique nous permettant le contrôle qualité des données ainsi que le calcul de différentes statistiques en fonction des données disponibles et du phénotype étudié.

L'haplotypage des données a été effectué à l'aide du logiciel Shape It [28] qui permet le phasage des chromosomes entiers. Le protocole de contrôle qualité utilisé est identique à celui de l'analyse et les options utilisées sont les options décrites dans la notice explicative présente sur le site internet (http://www.shapeit.fr/). L'imputation a été faite à l'aide du logiciel IMPUTE [67] sur les données haplotypées et des panels de références de 1000 génomes [6].

Néanmoins, PLINK ne permet pas l'analyse de données probabilistes comme celles issue de l'imputation. Nous utilisons alors le logiciel SNPTEST [103] qui nous permet alors d'effectuer les analyses standards sur les génotypes probabilistes. SNPTEST propose un score de qualité pour l'imputation ainsi qu'un score de qualité pour le calcul de la statistique pour chaque SNP testé. Nous avons utilisé un seuil de 90 % aussi bien pour la qualité de l'imputation que pour le calcul de la statistique. Ceci nous a permis préciser certaines régions d'intérêt en espérant trouver un polymorphisme (SNP/indel) présentant éventuellement plus d'intérêt plus évident.

Résultats

1. Genetropy

Genetropy est le logiciel que j'ai développé au cours de la thèse et qui permet de quantifier la part de données réellement indépendantes au sein d'un jeu de données génomiques, typiquement une cohorte génotypée à l'aide d'une puce de génotypage. La quantification de la part de redondance peut être utile notamment lors de la correction des tests multiples. Pour développer ce logiciel, j'ai utilisé l'entropie de Shannon qui permet de quantifier l'information ainsi que sa fraction indépendante.

Ce logiciel a fait l'objet d'une publication soumise au journal BMC Genomics qui est présentée comme résultat ci-après. Un résumé en français est aussi donné avant le texte de la publication.

a. Calcul par l'entropie de Shannon de la quantité d'information indépendante dans un jeu de données génomique

Résumé de la publication

A l'aide de travaux précédents dans lesquels l'entropie avait été utilisée comme mesure de déséquilibre de liaison [91, 92] et dans lesquels les auteurs ont démontré son aptitude à quantifier le déséquilibre de liaison (redondance d'information) aussi bien pour une paire de *loci*, mais aussi sur un plus grand nombre de *loci*, nous avons donc mis en œuvre un logiciel permettant d'estimer le nombre de SNPs indépendants au sein d'un jeu de données génomiques.

Pour ce faire, nous avons donc implémenté un phasage par paires de SNPs à l'aide d'un algorithme EM [25] afin de déterminer la redondance et ainsi donc la fraction indépendante de chaque paire de SNPs. Le logiciel se focalise alors sur une fenêtre de SNPs consécutifs dans laquelle toutes les paires de SNPs sont haplotypées et rassemble alors les SNPs de manière à ce qu'ils maximisent la redondance au sein de cette fenêtre de SNPs. Comme pendant à la redondance d'information, nous avons introduit une mesure de gain d'information. Elle correspond au rapport entre l'information de notre jeu de données sans redondance et l'information avec redondance.

Les résultats ont été encourageants, car avec un nombre minimal d'individus dans la population (n=60) et un nombre minimal de SNPs choisis dans la fenêtre de criblage (n=100), la quantité d'information indépendante calculée était stable et par conséquence le gain d'information aussi. Nous avons observé qu'un élargissement de la fenêtre de SNPs était accompagné d'un meilleur gain en information (stabilisé à partir de n=100 SNPs). Ceci s'explique par une plus grande fenêtre de SNPs et permet la capture de plus de déséquilibre de liaison (redondance d'information). De plus, le gain d'information est positivement corrélé avec la densité de SNPs présents au sein du jeu de données. Ceci s'explique par le fait qu'une grande partie des SNPs supplémentaires entre les différentes puces de génotypage sont redondants avec ceux déjà présents.

Nous avons calculé la quantité d'information indépendante pour plusieurs populations génotypées sur puces Illumina, les cohortes de 1000 génomes ainsi que pour des gènes

génotypés individuellement. Le gain d'information obtenu par notre logiciel varie entre 50 et 80%. A titre d'exemple, à partir des 21,5 millions de SNPs de la population européenne (381 individus), nous avons obtenu un gain d'information de 73,61% avec un nombre estimé de SNPs indépendant de 5,7 millions. Pour finir, nous avons testé une amélioration du logiciel en prenant en compte des données pré haplotypées et en adaptant l'algorithme, nous obtenons des résultats similaires ou meilleurs. Ces résultats préliminaires sont conformes aux propriétés de l'entropie de Shannon et sont encourageants pour le logiciel Genetropy.

Computation by entropy of the information contained in SNP datasets and applications

Lieng Taing[1], Cédric Coulonges[1], Sigrid Le Clerc[1], Sophie Limou[1], Christian Dina[2], Philippe Froguel[2], Matthieu Montes[1], Jean-Louis Spadoni[1], Hubert Cantalloube[1], Jean-François Zagury[1,*,§], Olivier Delaneau[1,*,§]

[1]Chaire de bioinformatique ; Laboratoire Génomique, Bioinformatique, et Applications

(EA4627), Conservatoire National des Arts et Métiers, Paris, France

[2]UMR CNRS 8090, Institut Pasteur de Lille, Lille, France

*These authors contributed equally to this work

§Corresponding authors

Email address:

JFZ: zagury@cnam.fr

OD: olivier.delaneau@gmail.com

Abstract

Background

We have developed a new method to determine the effective quantity of non redundant information stored in a genomic dataset (SNPs analyzed in a genetic region for a given population). This allows us to estimate the equivalent number of truly independent SNPs corresponding to this genomic dataset.

This method computes the entropy and uses the mutual information (MI), based on the linkage disequilibrium (LD) level between SNPs in the genetic region, to simply evaluate the effective non-redundant information in other words, an estimated number of independent SNPs contained in the genetic region. This method works at the genome level by including progressively each pair of SNPs that maximize the MI until covering all the elements of a defined window of SNPs.

Results

We developed a software, Genetropy, that computes the entropy and the MI in a genomic dataset. The redundancy of information depended mainly on the level of LD within the dataset, the population size, the genotyping density and on the SNP window size tested. Importantly, in genomic datasets from Caucasian (Europe) and Yoruba (Africa) ancestry, the redundancy reached an asymptote when dealing with population sizes larger than 75 or when using SNP windows larger than 200 SNPs. We thus applied this approach to compute the effective quantity of independent information in European genomic datasets derived respectively from the 300K and 1M Illumina beadchips and from the 21.5 M non monomorphic SNPs of the 1000 genomes Project (European population), and we obtained respectively 152,348, 289,393 and 5,667,633 estimated independent SNPs. Similar gain of information (50 to 80%) were also observed at the level of single genotyped genes. The

running time of the program was very rapid, for instance less than 10 minutes for 350 subjects genotyped in a 300K SNP chip.

Conclusion

At a time of large-scale sequencing/genotyping of populations, this approach provides a simple and precise description of the information and redundancy (i.e. LD level) found in genomic datasets. This approach could also help refine the Bonferroni significance threshold by providing the true number of independent tests performed on the dataset, but in our experience the impact was minor.

Background

In the past few years, the availability of powerful genotyping chips has allowed the completion of numerous genome-wide association studies (GWAS). These chips rely both on the technical progress and on the increased knowledge gained from the HapMap project [1-3] and more recently from the 1000 genomes project [4]. The number of SNPs in one chip has increased from 120,000 in 2003 to 2.5 million today (new Illumina technology) expanding the resolution by a factor 10 within a few years. With the progress of sequencing technology, it is likely that GWAS will soon handle most of the human genetic variations, in particular the 40+ million suspected SNPs of the genome [4].

Concomitant to these technological achievements, statistical issues linked to multi-testing have arisen. Scientists generally use a 5% false positive p value threshold when making a statistical test, in other words, if the probability (p value) of observing the tested conformation is below the 5% threshold, the H0 hypothesis (no association between a SNP and the tested phenotype) is rejected and there is an association. However, this 5% threshold must be adapted for genotyping chip analysis since for 1 million SNPs, one could expect 50,000 SNPs exhibiting a p value below 5% due to the uniform distribution of the p values, and thus declare positive results even if they are false. To alleviate the lack of confidence caused by multi-testing, several methods have been introduced to correct the p-values. The Bonferroni correction, based on the assumption that all the tests are independent, is the most widely used and divides the raw p values by the number of tests performed. This approximation is however over-conservative and not always adapted for a regular GWAS.

For instance, for a GWAS on one million SNPs, the threshold becomes 5×10^{-8}. It will be very difficult to match this threshold for any study unless it involves thousands of cases and controls, or the impact of the tested gene variant on the phenotype is strong (typically odds ratio greater than 3).

Alternative methods have been developed to assess the proportion of false-positives among the rejected null hypotheses. For instance, the False Discovery Rate [5] assesses the proportion of false positive signals among the N best p values of a study, or the random permutation of patient labels looks for empirical p value distributions [6]. In all these corrections, one assumes that the SNPs are independent.

The assumption of SNP independency is key in the issue of over-correcting the p values. SNPs in the genome are usually locally correlated [7-8] and some SNPs may exhibit a high level of linkage disequilibrium (LD measured by the R^2 coefficient) leading to correct the type I error for 2 or more tests when one should not. The inflation on the corrected threshold also occurs when the correlation between SNPs is only partial, consequently diminishing the power to detect true signals.

Manufacturers of genotyping chips have tried to limit this issue by selecting only representative SNPs called Tag SNPs and also by maximizing the coverage of the genome thanks to the HapMap and 1000 genomes projects. The use of Tag SNPs has increased the power of genotyping chips, but numerous SNPs still exhibit a substantial correlation (r^2>0.7). Several computational methods can be applied to take into account the correlation between SNPs. The Principal Component Analysis, widely used in data analysis, can help identify several axes of independent SNPs [9]. Alternatively re-computation of really independent Tag SNPs for each genotyping chip could also be considered. A recent study has compared these approaches which appear rather time-consuming [10]. Another study has put forward a universal p value threshold of 7.2 x 10-8 for standard GWAS [11], however it did not take into account the surprisingly high density of SNPs currently depicted by the 1000 genomes project [4]

In the present work, we have developed an original method to rapidly and precisely compute the amount of non redundant information in a genomic dataset yielding an estimated

number of equivalent independent SNPs that should help apprehend the issue of over-conservation. Its principle is 1. To compute an entropy for the dataset -equivalent amount of information bits (the bit is the elementary unit for information) corresponding to the SNPs-, 2. To determine the possible compression of information thanks to the mutual information computed through the knowledge of the LD between SNPs, 3. One can then estimate the equivalent number of independent SNPs in the genomic dataset by a simple rule of three and the corresponding gain of information can also be computed.

6

<

Implementation

Rationale and algorithm

Let be a group of SNPs genotyped in a population over a given genomic region (a gene, a chromosome fragment, or the whole genome). The idea of the method is to create a minimal set of SNPs spanning the maximum of information, by adding the SNPs progressively one by one and computing the real information according to the level of LD, in order to get finally the real independent information derived from all these SNPs. By comparing this compacted information with the entropy cumulated from each individual SNP, the level of compression of information will be found. From the initial number of SNPs, it will thus be possible to estimate the equivalent number of independent SNPs for the genomic region by a simple rule of 3. We present hereafter the 6 detailed steps to perform this computation.

1. Entropy of a single SNP.

Considering a SNP A with two alleles a1 and a2 with allelic frequencies fa1 and fa2, its entropy (information) can be computed from its allelic frequencies:

$$E(A) = -f_{a_1} \log_2 f_{a_1} - f_{a_2} \log_2 f_{a_2}$$

For a SNP with a 0.5 minor allele frequency (MAF), the entropy value will be maximal with a value of 1 bit. As the MAF decreases, the entropy will decrease to reach 0 for a monomorphic SNP.

2. Pair wise SNP haplotyping

The aim of this step is to infer haplotypes in order to measure the linkage disequilibrium between two SNPs. We used the rapid well-known Expectation-Maximisation algorithm [12]. It is not the most accurate haplotype inference method but as shown in the results section, it provides satisfying results while being extremely fast.

3. Computation of SNP pair entropy

Let's consider two SNPs A, with two alleles a1 and a2, and B, with two alleles b1 and b2. Through step 2, we can determine the four possible haplotypes, and the haplotype entropy associated with these 2 SNPs as a locus will be as follows:

$$\begin{aligned} E(A\_B) = & -f_{a_1b_1}\log_2 f_{a_1b_1} \\ & -f_{a_1b_2}\log_2 f_{a_1b_2} \\ & -f_{a_2b_1}\log_2 f_{a_2b_1} \\ & -f_{a_2b_2}\log_2 f_{a_2b_2} \end{aligned}$$

The entropy of the haplotype is maximum when the two SNPs are fully independent ($f_{a_ib_j} = f_{a_i} \times f_{b_j}$ for all i and j) and in that case, its value is the sum of the entropy of each SNP.

4. Computation of the mutual information between 2 SNPs.

If two SNPs are not fully independent we can use the mutual information (MI) as a measure of the correlation between the two SNPs [13-14] as follows:

$$MIR = E(A) + E(B) - E(A\_B)$$

Interestingly, when dealing with two independent SNPs (D=0), the MI will be equal to 0 whatever their respective frequencies.

5. SNP pair aggregation based on the mutual information, using a Kruskal algorithm.

As seen in steps 1 to 4, it is possible to compute the entropy of each SNP, of each haplotype of two SNPs, as well as their MI. A genomic region can be then represented by a graph defined as follows:

- each node is a SNP
- Each possible pair of nodes are connected by edge whose weight is their MI (measure of their correlation)

The method uses a Kruskal algorithm [15] to cover optimally all the SNPs, so that they can be all connected by a path. This algorithm provides a set of non cyclic edges that regroup all the SNPs in the region and that maximise the correlation between each SNP (i.e.

maximisation of the mutual information), this set being called the maximum spanning tree (MST). The principle of the Kruskal algorithm is reminded in Figure 1.

6. Computation of the independent information and of an estimated number of independent SNPs in the region.

In our case we can compute the number of independent bits for the genomic region by generalizing the MI formula:

$$I(genomic\ region) = \sum E(SNPs) - E(MST)$$

By construction this formula provides the number of independent bits necessary to capture the genomic region information. We can thus determine a ratio of redundancy and estimate an equivalent number of independent SNPs as follows:

$$Estimated\ independent\ SNPs = Nb\ SNPs \times \frac{I(genomic\ region)}{E(SNPs)}$$

The ratio between the independent information of the genomic region and its entropy measures the information redundancy (1 = no redundancy, $\frac{1}{Nb\ SNPs}$ = full redundancy). This estimated number of independent SNPs could be used for the multitesting corrections.

7. Possible generalization of our approach.

Interestingly; this approach could be generalized to larger haplotypes (three and more SNPs) and also to multi-allelic polymorphisms (three alleles and more). This could be done simply by modifying the mode of aggregation in the Kruskal algorithm, but it requires pre-phased data (data not shown). This should be more precise than limiting oneself to SNP pairs, however the pre-phasing step is too restrictive. An example of the results is provided in table 3 for single genes. We could not use this approach at the whole genome level since we did not have the pre-phased data at the whole genome level.

Adaptation for genomic data

In order to program the previous algorithm, we need to work with a SNP window that could range from the size of a single gene (a few SNPs) to a whole chromosome (thousands of SNPs). The latter approach is not adapted since 1) most of the LD (i.e. loss of information) is generally contained in neighbouring SNPs within the 250 kb range and 2) the matrix size would heavily slow down the computation speed for the haplotype reconstruction. As an example, there are 3.5 M SNPs in the sequenced chromosome 2 of the 1000 genomes project, and we would have to compute and store about 6.2×10^{12} possible edges.

For that reason, a window of reasonable size is better suited for genome-wide data. As shown in the Results section, window sizes of 100 to 500 consecutive SNPs are easy to deal with. In a one million SNPs genotyping chip, the space between 2 consecutive SNPs is in average 3 kb, and 300 SNPs would thus span over a 1 Mb region, which is generally large enough to include most SNPs with a significant LD level. In order to take into account the LD between SNPs close to the boundary of two neighbouring windows, we will simply need to perform a second pass over the chromosome, shifted by half a window size. The number of estimated independent SNPs in the chromosome is then simply obtained as the mean number between the two passes.

Material and Methods

To assess the overall linkage disequilibrium between SNPs, the speed and the stability of our method, we used two types of genomic datasets: 1. genotypes obtained by high throughput genotyping (Illumina, Affymetrix) from several cohorts, 2. haplotypes datasets obtained by sequencing from 1KGP for several ethnic groups especially several genes in the GRIV cohort[16-18].

The GRIV cohort

The GRIV (Genomics of Resistance to Immunodeficiency Virus) cohort has been established in France in 1995 to generate a large collection of DNA for genetic association

studies to identify host genes associated with extremes response to AIDS. Only subjects of European descent living in France were eligible for enrolment to reduce confounding effects linked to population stratification. The cohort is composed of 359 subjects and GWAS results obtained from Illumina HumanHap300 beadchips have been previously published [19-21]. All the patients gave their informed consent and the study was approved by the Institutional Review Board of Saint-Louis hospital (Paris, Fance).

The DESIR cohort

The D.E.S.I.R. cohort (Data from an Epidemiological Study on Insulin Resistance Syndrome) is a cohort of 697 non-obese, normoglycemic, HIV-1 seronegative French subjects recruited from 1994 to 1996 as negative controls for the D.E.S.I.R. trial [22]. This cohort was genotyped with Illumina HumanHap 300 beadchips. All the patients gave their informed consent and the study was approved by the Institutional Review Board of Kremlin-Bicêtre hopsital (Kremlin-Bicêtre, France).

The CTR cohort

A cohort of healthy Caucasian subjects (n=502) was genotyped on the 1M Illumina Omni quad beadchips to serve as control in a GWAS and these genotypes were tested with our method. All the patients gave their informed consent and the study was approved by the Institutional Review Board of Cochin hospital (Paris, France) [23].

The 1000 Genomes Project

The 1000 Genomes Project was launched in 2008 by an international research consortium with the aim of creating a new map of the human genome with a increased resolution [4]. The June 2011 release of the 1000 Genomes Project has yielded the genotypes of 1094 individuals for 37,426,733 SNPs including singletons (only one variant allele observed among the whole population). We have focused our analysis on the population of European descent composed of CEU (87 subjects Utah resident of Northern and Western

European ancestry), FIN (93 Finnish subjects from Finland), GBR (89 British subjects from England and Scotland), IBS (14 Iberian subjects from Spain) and TSI (98 Toscani subjects from Italia) for a total of 381 subjects of European ancestry covering 21,688,871 non-monomorphic SNPs. We also used the YRI population (98 Yoruba subjects from Kenya) to assess the behaviour of our method on a non Caucasian population.

Quality Control of the genotyping chips

In spite of the standard quality controls set in both Affymetrix and Illumina platforms, some genotyping errors may remain. Additional filters were thus used for quality control at the level of both SNPs and individuals. The MAF threshold used in the present work was set at 5%. The p value threshold for the Hardy Weinberg equilibrium test to keep a SNP genotyped in the population was set at 10^{-3}. For each genotyped subject, the threshold for missing data was set at 2%. The 1000 Genomes Project data are not filtered since they are used as reference data for the haplotyping validation.

Validation of haplotyping

Out first priority was to validate the EM algorithm that we used in our method to infer the haplotypes, since this step is critical for the evaluation of the LD between SNPs. We used the genotyped data from the 60 subjects of the 1000 Genomes Project, focusing on chromosome 1. We sliced chromosome 1 into windows of 300 SNPs and for each pair of SNPs within a window, we compared the haplotype frequencies obtained by the reference software Phase 2.1 [24] with the frequencies obtained by SNPs base pair haplotype EM calling. We defined a score based on the absolute value of the difference between our EM and Phase 2.1 4 haplotype frequencies and computed the mean of this score among the SNP pairs derived from the 300 SNP-windows, covering the entire chromosome.

$$\Delta_{SNP\ pair\ haplotype} = \sum_{h=1}^{h=n} \left| F_{Observed}(h) - F_{Computed}(h) \right|$$

n corresponding to the four possible haplotypes in a pair of SNPs.

Computation of the estimated number of independent SNPs and gain of information obtained through the Genetropy software.

As shown in step 6 of the algorithm, the estimated number of independent SNPs was computed as follows:

$$Estimated\ Independent\ SNPs = Number\ of\ SNPs \times \frac{Independent\ Bits}{Initial\ \ Equivalent\ Bits}$$

We also defined the gain of information (%) obtained when comparing the equivalent number of bits (obtained by computing the entropy of the genomic region) and the number of independent bits (obtained after aggregating SNPs and computing MI) by the following formula:

$$Gain\ of\ information = \left(1 - \frac{Independent\ Bits}{Initial\ Equivalent\ Bits}\right) \times 100$$

This simple measure enabled us to compare data from different cohorts genotyped on various platforms.

Tests performed for the measure of the SNP entropy

We first assessed the haplotyping quality. We then evaluated the impact the SNP window size which is the main parameter of the entropy- computing program. We ran our method on three different samples, the whole GRIV cohort, the whole CTR cohort, and the Chromosome 2 of the 1000 Genomes project for different SNPs window sizes ranging from 25 to 300 SNPs by 25 SNPs increment. We then investigated on the impact of population size: random samples of size ranging from 25 to 600 subjects (with 25 subject increments) were extracted from the GRIV, DESIR, and CTR and tested. We also tested the impact of population size in the 1000 Genomes EUR population by extracting sample populations with sizes ranging from 10 to 380 subjects (with 10 subject increments). We then investigated if ethnicity or some biological phenotypes could influence the result of our program. For that, we compared the gain of information obtained in the 1000 Genomes CEU population, the

1000 Genomes YRI project and the merged cohorts, and we also compared 2 Caucasian cohorts, the GRIV cohort and the DESIR control cohort [19-21].

Results

Haplotyping quality

The comparison of the SNP pair haplotype frequencies obtained by the EM algorithm and by the reference software Phase 2.1 from the 1000 Genomes project data yielded an average absolute error of 1 to 2 % over the whole chromosome 1. This difference appears negligible for the computation of linkage disequilibrium over a genetic region and confirms that the rapid EM algorithm should be relevant for the entropy computation.

As discussed in the Methods, the quantity of information contained in various sets of SNPs may depend on several parameters: window size, population size, SNPs density in the region, ethnicity and phenotype. We thus explored the effect of these parameters. In the following, the results are presented indifferently either under the form of an estimated number of independent SNPs or under the form of a gain compared to the initial number of genotyped SNPs, since it is equivalent.

Window size effect

We first studied the effect of the SNP window size, the main parameter of the entropy-computing program, on the gain of information. We tested several window sizes ranging from 10 to 300 consecutive SNPs covering the 22 autosomes in the GRIV (n=260) and DESIR (n=697) cohorts genotyped with the Illumina 300k beadchip and in the CTR cohort (n=502) genotyped with the Illumina 1M beadchip, or covering the chromosome 2 in the 381 European subjects from the 1000 Genomes project. For the genomic chip data, we observed a similar curve with a quasi-plateau reached after 50 SNPs (Figure 2A). For the 1000 genome data, the density of SNPs (in average 1 SNP every 300 bp) makes the curve less asymptotic since you may still get LD information at 300 SNPs (corresponding to a 100 kb span), however the curve gets much flatter after 200 SNPs.

The gain of information is very slightly increasing with larger SNP window sizes, but the increase of 1% observed between windows of 100 and 400 SNPs for sample size bigger than 75 individuals in the GRIV data is associated with a much longer running time (Table 1). In the following, we have thus used a window size of 100 SNPs.

Population size effect

To determine the impact of population size on the gain of information, we took random subgroups of various sizes from the GRIV, DESIR, CTR cohorts testing subgroups ranging from 25 up to 700 subjects. We observed that the gain decreased for growing population sizes and reached a plateau for subgroups larger than 60-75 individuals (Figure 2B). We also performed the same analysis on the chromosome 2 of the 1000 Genomes project European population for different sizes ranging from 10 to 380 people by increments of 10 individuals (each size was sampled 10 times).

Overall, Table 1 shows that the population size may have slightly more impact on the gain of information than the SNPs window size, however this impact remains minor (about 2% between 60 and 200 subjects).

Effect of the population origin and phenotype

We looked at the impact of the origin of the cohort by comparing the gains of information between European cohorts such as Finish, British, Toscan, CEU, and Iberian populations. They exhibited a similar behaviour with a similar gain of information of 80 % (data not shown). For instance, for the 87 CEU with 11,16 M non monomorphic SNPs, the gain observed (about 80%) leads to 2,160,617 estimated independent SNPs. When merging the 1000 Genomes project European population (n=381) the gain of information was lower at 73% yielding 5,667,633 estimated independent SNPs from the 21,47 M initial SNPs. This lower gain likely results from the diversity of Finish, British, Toscan, CEU, and Iberian

populations, reflecting the increase in singletons specific of each origin. Interestingly, the African YRI population (n=65) also exhibited an 80% gain of information.

We also looked at the impact of a particular phenotype on our results by comparing the gain obtained in HIV-1 infected subjects from the GRIV cohort and in control subjects from the D.E.S.I.R. cohort. These groups are of European descent and were analyzed in GWAS to look for genetic factors involved in AIDS progression or infection [19-21]. There was hardly any difference between these two groups (Figure 2A and 2B). This can be explained by the fact that the potential differences are localized on relatively few SNPs and their impact is minor over the whole genome.

Impact of the SNP density

The entropy computation involves the level of LD among the SNPs. When the SNP density is high (i.e. suggesting more LD among the SNPs) there should be a global decrease of information relatively to the number of SNPs. When the SNP density is low (i.e. suggesting a lower LD) there should be a global increase of information relatively to the number of SNPs. This is indeed what is observed in Figure 2A where the gain of information is maximum for the very dense 1K genome data, then for the OmniOne beadchip (1M SNPs), then for the 300 K Illumina beadchip.

For the 300K Illumina beadchip we found 152,430 estimated independent SNPs for the DESIR cohort (47% gain of information) and 152,347 estimated SNPs for the GRIV cohort (47% gain of information). For the Illumina 1M OmniOne beadchip we found 289,393 independent SNPs (63% gain of information). For the sequenced data of 1KGP we found 2,160,617 estimated independent SNPs (80.6 % gain of information) for the CEU cohort (N = 87) from the 11,16 M known non monomorphic SNPs. These observations confirm the expected positive correlation between the density of SNPs in the genotyped data and the gain of information.

Impact of the SNP minor allele frequencies (MAF)

We also looked at the impact of the considered SNP MAFs in the European population from 1KGP (Figure 3). We observed a similar gain of 83% with a MAF of 5% (going from 6,444,254 SNPs to 1,095,729 estimated independent SNPs) and a MAF of 10% (going from 5,091,085 SNPs to 830,657 estimated independent SNPs). Since singletons represent 9 % of the 1KGP SNPs for the whole European populations (N = 381) and are by essence independent, the gain was lower when considering all the SNPs (73%).

Analysis at the gene-level

Since many published works deal with a single gene (candidate gene studies), it was also interesting to evaluate the estimated number of independent SNPs in localized gene regions. We tested candidate genes previously analyzed in the GRIV cohort [16-18]. Table 3 presents the gain of information and estimated number of independent SNPs for those genes.

We have also tested the generalized method based on the data directly phased over the whole gene (this method could not be implemented over the whole genome due to complexity issue, see the algorithm description) and recomputed the Estimated Independent SNPs from the genotypes of 1KGP. The results in Table 3 show that the generalized method yields a better compression, but it is not so much better than the method using the simple SNP pair aggregation.

Running time

Table 1 presents the running time and gain obtained in the GRIV cohort according to the population sample size and according to the window size chosen. Logically, the running time increases with the sample size and the window size. The gains became rather stable with a window size starting at 100 and a population size at 60-75 (Table 1). For 300,000 SNPs, the gain could be computed in 4 mn.

For the 87 CEU subjects of the 1000 Genomes project, the running time was also quite rapid with 50 minutes for the whole genome.

Discussion and Conclusion

We have presented a new methodology based on entropy and mutual information to estimate the equivalent number of independent SNPs in a genomic dataset and a putative gain of information. The gain of information depends on very few parameters: the number of genotypes, the number and density of SNPs, and the window size chosen to analyze the dataset and to a lesser extent on the ethnicity of the population. Overall, the gain of information reached a plateau for populations larger than 60-75 individuals, it also reached a plateau when screening with window sizes larger than 100 SNPs (i.e. corresponding to a span of 500kB in 300K chips). One must be cautious regarding the size of the cohort since the haplotype frequencies can be indeed biased when dealing with less than 30 subjects. An important observation was the higher gain of information obtained for a higher SNP density (Table 2). This is simply explained by the fact that the SNPs added to increase the density are more likely to be in LD with already present SNPs. This observation constitutes an additional evidence that the diversity among human genomes may reach a limit and this is in agreement with the idea of using a universal p value threshold for GWAS [11].

The gain of information was slightly similar among Caucasian populations but a bias was introduced when combining them due to singleton SNPs. Interestingly, the gain was similar for the African YRI group.

In light of this gain of information, we reanalyzed the data published in our previous AIDS GWAS [20-22] using the newly estimated number of independent SNPs for the 300K Genomic beadchip (Table 2), but it did not reveal any new association (data not shown).

Finally the MI has proven to be a relevant measure of linkage disequilibrium [13-14] and could be generalized for multi-allelic polymorphisms (such as haplotypes or copy number variation) or for multiple locus polymorphism. We indeed computed the gain of

information at the gene level using a generalized haplotype-haplotype aggregation method and preliminary results showed a slight increase in precision.

In conclusion, we have presented a new and rapid way to compute independent information derived from SNP datasets and the results found were robust for large enough population and large enough screening SNP windows. The software is available to the scientific community and it should be a useful tool for genomic applications. Beyond the estimation of the number of independent SNPs, one can foresee the rapid characterization of the LD level of a genomic region as an important application of this approach.

Availability and requirements

Project name: Genetropy

Project home page: http://griv.org/genetropy

Operating system(s): Unix (Ubuntu)

Programming language: c++

Other requirements: none

License: Free software licence

Any restrictions to use by non-academics: none

List of abbreviations

LD: Linkage Disequilibrium

GWAS: Genome Wide Association Studies

1KGP: 1000 genomes project

MAF: Minor allelic frequency

Competing interest

The authors have no conflict of interest to declare.

Authors' contributions

LT worked on developing the methods and programs used in this study under the direct supervision of both JFZ and HC who conceived the study. OD provided advice and assistance

for the algorithm and software implementation. OD, CC, SLC, SL, PF, CD, MM and JLS provided genomic data. All the authors have read and approved the final manuscript.

Acknowledgements

The authors are grateful to the individuals who kindly provided their DNA for genetic studies. The project was supported by a grant from Conservatoire National des Arts et Métiers and from CERIES. LT was funded by The French Ministry of Education and Research, OD was funded by a grant from Agence Nationale de Recherches sur le SIDA (ANRS).

References

[1] The International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q et al.: **A second generation human haplotype map of over 3.1 million snps**. *Nature* 2007, **449**: 851-861.

[2] The International HapMap Consortium: **A haplotype map of the human genome**. *Nature* 2005, **437**: 1299-1320.

[3] The International HapMap Consortium: **The international hapmap project**. *Nature* 2003, **426**: 789-796.

[4] 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**: 1061-1073.

[5] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society, Series B* 1995, **57 (1)**: 289–300.

[6] Browning BL: **Presto: rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one and two-stage genetic association studies**. *BMC Bioinformatics* 2008, **9**: 309.

[7] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: **High-resolution haplotype structure in the human genome**. *Nat Genet* 2001, **29**: 229-232.

[8] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome**. *Science* 2002, **296**: 2225-2229.

[9] Gao X, Starmer J, Martin ER: **A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms**. *Genet Epidemiol* 2008, **32**: 361-369.

[10] Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ: **Accounting for multiple comparisons in a genome-wide association study (gwas)**. *BMC Genomics* 2010, **11**: 724.

[11] Dudbridge F, Gusnanto A: **Estimation of significance thresholds for genomewide association scans**. *Genet Epidemiol* 2008, **32**: 227-234.

[12] Fallin D, Schork NJ: **Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data**. *Am J Hum Genet* 2000, **67**: 947-959.

[13] Nothnagel M, Fürst R, Rohde K: **Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks**. *Hum Hered* 2002, **54**: 186-198.

[14] Zhang L, Liu J, Deng H: **A multilocus linkage disequilibrium measure based on mutual information theory and its applications**. *Genetica* 2009, **137**: 355-364.

[15] Kruskal JBJ: **On the shortest spanning subtree of a graph and the traveling salesman problem**. *Proc. Amer. Math. Soc.* 1956, **7**: 48-50.

[16] Do H, Vasilescu A, Carpentier W, Meyer L, Diop G, Hirtzig T, Coulonges C, Labib T, Spadoni J, Therwath A, Lathrop M, Matsuda F, Zagury J: **Exhaustive genotyping of the interleukin-1 family genes and associations with aids progression in a french cohort**. *J Infect Dis* 2006, **194**: 1492-1504.

[17] Do H, Vasilescu A, Diop G, Hirtzig T, Coulonges C, Labib T, Heath SC, Spadoni J, Therwath A, Lathrop M, Matsuda F, Zagury J: **Associations of the il2ralpha, il4ralpha, il10ralpha, and ifn (gamma) r1 cytokine receptor genes with aids progression in a french aids cohort**. *Immunogenetics* 2006, **58**: 89-98.

[18] Diop G, Hirtzig T, Do H, Coulonges C, Vasilescu A, Labib T, Spadoni J, Therwath A, Lathrop M, Matsuda F, Zagury J: **Exhaustive genotyping of the interferon alpha receptor 1 (ifnar1) gene and association of an ifnar1 protein variant with aids progression or susceptibility to hiv-1 infection in a french aids cohort**. *Biomed Pharmacother* 2006, **60**: 569-577.

[19] Limou S, Coulonges C, Foglio M, Heath S, Diop G, Leclerc S, Hirtzig T, Spadoni J, Therwath A, Lambeau G, Gut I, Zagury J: **Exploration of associations between phospholipase a2 gene family polymorphisms and aids progression using the snplex method**. *Biomed Pharmacother* 2008, **62**: 31-40.

[20] Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, Delaneau O, Labib T, Taing L, Sladek R, Deveau C, Ratsimandresy R, Montes M, Spadoni J, Lelièvre J, Lévy Y, Therwath A, Schächter F, Matsuda F, Gut I, Froguel P, Delfraissy J, Hercberg S, Zagury J: **Genomewide association study of an aids-nonprogression cohort emphasizes the role played by hla genes (anrs genomewide association study 02)**. *J Infect Dis* 2009, **199**: 419-426.

[21] Le Clerc S, Limou S, Coulonges C, Carpentier W, Dina C, Taing L, Delaneau O, Labib T, Sladek R, , Deveau C, Guillemain H, Ratsimandresy R, Montes M, Spadoni J, Therwath A, Schächter F, Matsuda F, Gut I, Lelièvre J, Lévy Y, Froguel P, Delfraissy J, Hercberg S, Zagury J: **Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for aids (anrs genomewide association study 03)**. *J Infect Dis* 2009, **200**: 1194-1201.

[22] Balkau B: **[an epidemiologic survey from a network of french health examination centres, (d.e.s.i.r.): epidemiologic data on the insulin resistance syndrome]**. *Rev Epidemiol Sante Publique* 1996, **44**: 373-375.

[23] Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, Roussel A, Favier A, Briançon S: **The su.vi.max study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals**. *Arch Intern Med* 2004, **164**: 2335-2342.

[24] Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *American Journal of Human Genetics* 2001, **68**: 978-989.

Tables

Sample Size	SNP Window Size: 100		200		300		400	
30	3,9	51,1%	8,7	52,6%	12	53,1%	14	53,3%
40	4,2	50,6%	8,8	51,1%	13	51,4%	16	51,3%
60	5,1	49,2%	9,8	49,5%	16	50,2%	17	50,2%
75	5,9	48,5%	10	49,2%	18	49,5%	18	49,6%
150	6,3	47,7%	15	48,3%	22	48,5%	24	48,3%
225	8,4	47,1%	18	47,7%	26	47,9%	30	48,0%
300	9,3	47,4%	18	47,7%	26	47,9%	35	48,0%

Table 1 — Computation time (mn) and gain (%) obtained according to the sample size and the window size on the GRIV cohort

Description of gain of information and computation time on the G.R.I.V. cohort dataset (300K Illumina beadchip) for several sample and window sizes.

Technology	Number of SNPs	Estimated number of Independent SNPs
Illu. 300k (GRIV)	287.919	152.347
Illu. 300k (DESIR)	287.894	152.429
Illu. 1M (CTR)	788.778	289.393
1K Genomes Project CEU	11,160,516	2,160,617
1K Genomes Project YRI	11,181,516	2,158,705
1K Genomes Project EUR	21,474,139	5,667,633

Table 2 – Estimated Independent SNPs obtained in several common genetic datasets

Estimated independent SNPs computed with one hundred SNPs window size, based on our several datasets.

GENE	SNPS	Estimated Independent SNP regular method	Estimated Independent SNP generalized method
TNFRSF1A	1	1	1
FASLG	2	1.6315	1.67
IFNG	2	1.48929	1.48974
IFNGR1	2	1.98343	1.97
IL1B	3	2.6919	2.59
IL4	3	1.66936	1.66911
ZAP70	3	2.07055	2.06
IFNAR1	4	3.10817	3.02
LTA	4	3.13422	3.13
CXCR2	5	2.62059	2.55
GPR15	5	3.90258	3.82
CXCR6	6	4.10998	3.93
FAS	6	5.12038	4.67
IL10RA	6	4.2659	3.55
TNF	7	4.6207	4.0025
IL6	8	4.16706	3.7
IL10	9	2.93956	2.90737
IL1RN	9	3.60723	3.503
IL1A	10	2.75727	2.717
TRIM5	10	6.47223	5.3

Tableau 3 – Application to interest genes with regular method and generalized method
Application of our method on several gene datasets with a classical approach on genotype data and with a generalized methodology on phased data. The classical approach infers SNP pair haplotypes using EM and gathers SNPs by pairs; the generalized method gathers the SNPs progressively beginning with simple pairs up to one block recovering all the dataset, using the pre-phased data (see Methods).

Figures

Figure 1 – Illustration of the aggregation process by the Kruskal's algorithm.
The scheme describes the behaviour of the Kruskal's algorithm for a given set of SNPs. The aim of the algorithm is to determine the Maximal Spanning Tree that allows to maximize the linbkage disequilibrium among a set of SNPs, which corresponds to the weight marked on each edge. The algorithm progressively merges sets of SNPs exhibiting the highest LD found between them, until there is only one set of SNPs. The tree which has been produced through this process has no cycles.

In step A: Computation of mutual information between all possible pairs of SNPs. The dotted lines have been removed because they represent full independency (MI=0) between two SNPs. Each SNP constitute its own set of SNPs.

In step B: The pair of SNPs exhbiting the highest LD is merged into one set. Here, there is equality between two pairs of SNPs, and one is chosen arbitrarily.

In step D: The pair of SNPs with highest LD (SNP 5 - SNP 6) includes an already merged SNP (SNP 5). Then the SNP 6 is merged with the previously built set of SNPs (SNP 1-SNP 5).

In step E: The following pair of SNPs with highest LD (SNP 1 - SNP 2) is merged. The algorithm then bans the cyclic edges (SNP 5 - SNP 2).

In step F: The algorithm bans the cyclic edges (SNPs 2-3; 4-5; 4-6).

In step G: The Maximum Spanning Tree recovers all the SNPs of the graph. There is only one set of SNPs and the algorithm stops.

Figure 2 - Assessing the gain of information by SNP window size or by sample size.
The gain of information was computed in the GRIV cohort (n=260) and the DESIR cohort (n=697) genotyped on the 300K Illumina beadchip, on the CTR cohort (n=502) genotyped on the 1M OmniOne Illumina beadchip, and on the chromosome 2 from the 1000 Genomes

project corresponded to 1,863,376 SNPs in 381 European subjects (see Methods) The blue and green curves for GRIV and DESIR are completely overlapping.

A. effect of various SNP window sizes. Each computation was made on the whole cohort.

B. effect of various sample sizes. Each sample size was tested ten times and only the mean value was presented, the square deviation was less than 1% and is not shown.

Figure 3 - Assessing the gain of information on European panel using different MAF threshold

The gain of information was computed on 381 European genotypes from 1KGP and the numbers of SNPs and estimated independent SNPs are presented according to the MAF (>0%, >5%, >10%). The initial number of SNPs is in deep blue, the number of estimated independent SNPs is in light blue.

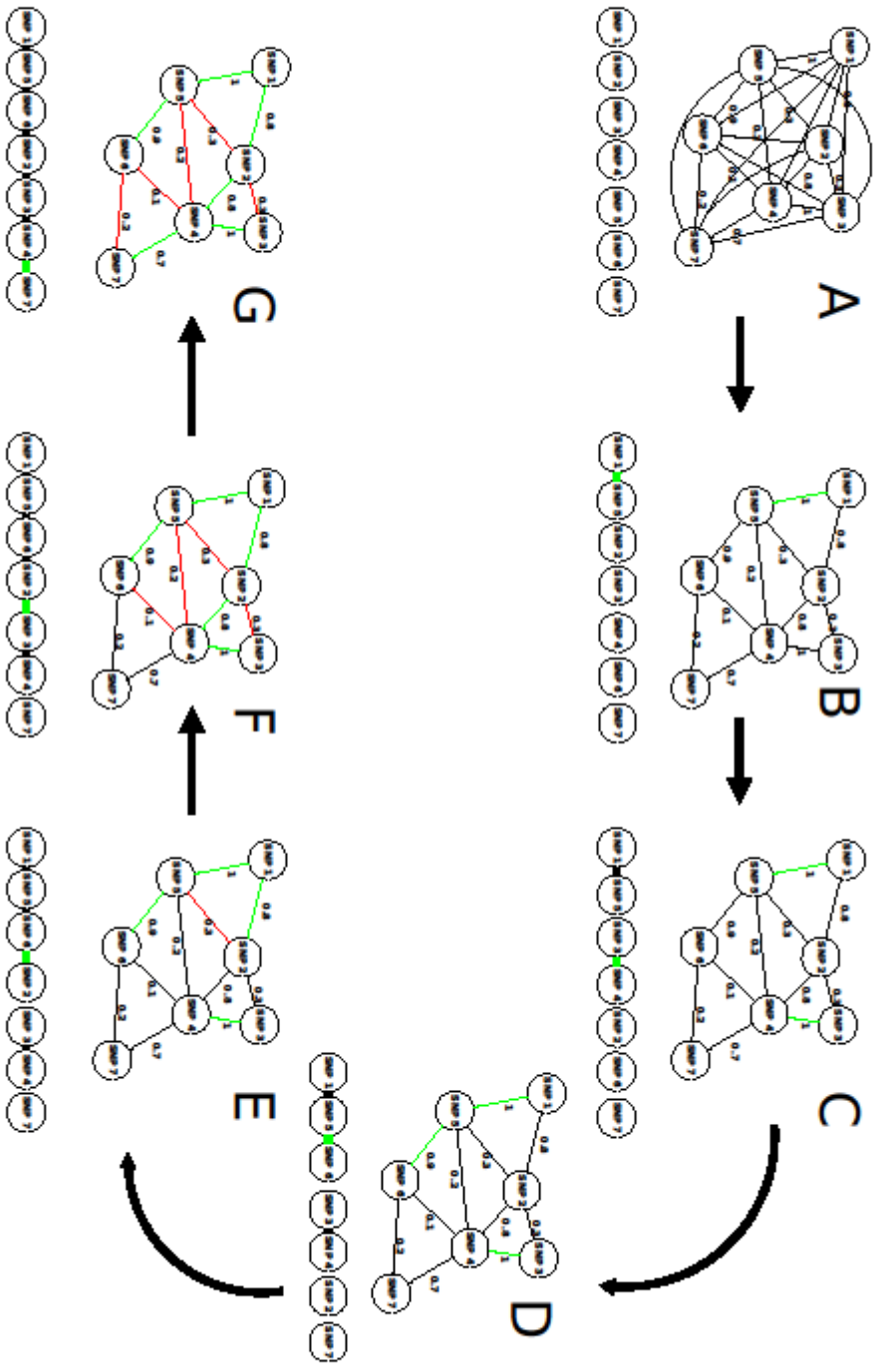

Figure 1

Référencé comme la figure 1 dans le manuscrit de la publication.

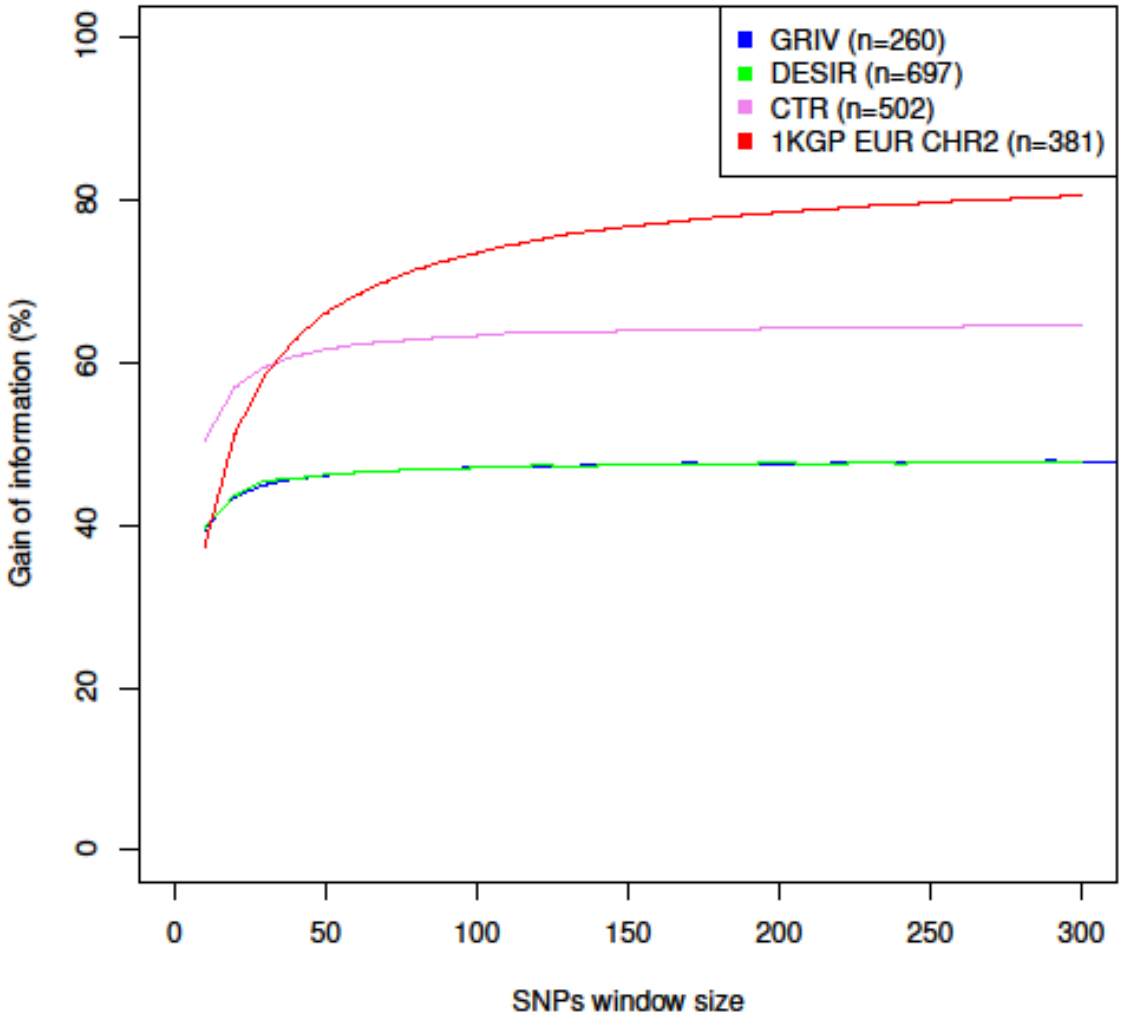

Figure 2

Référencé comme la figure 2.A dans le manuscrit de la publication.

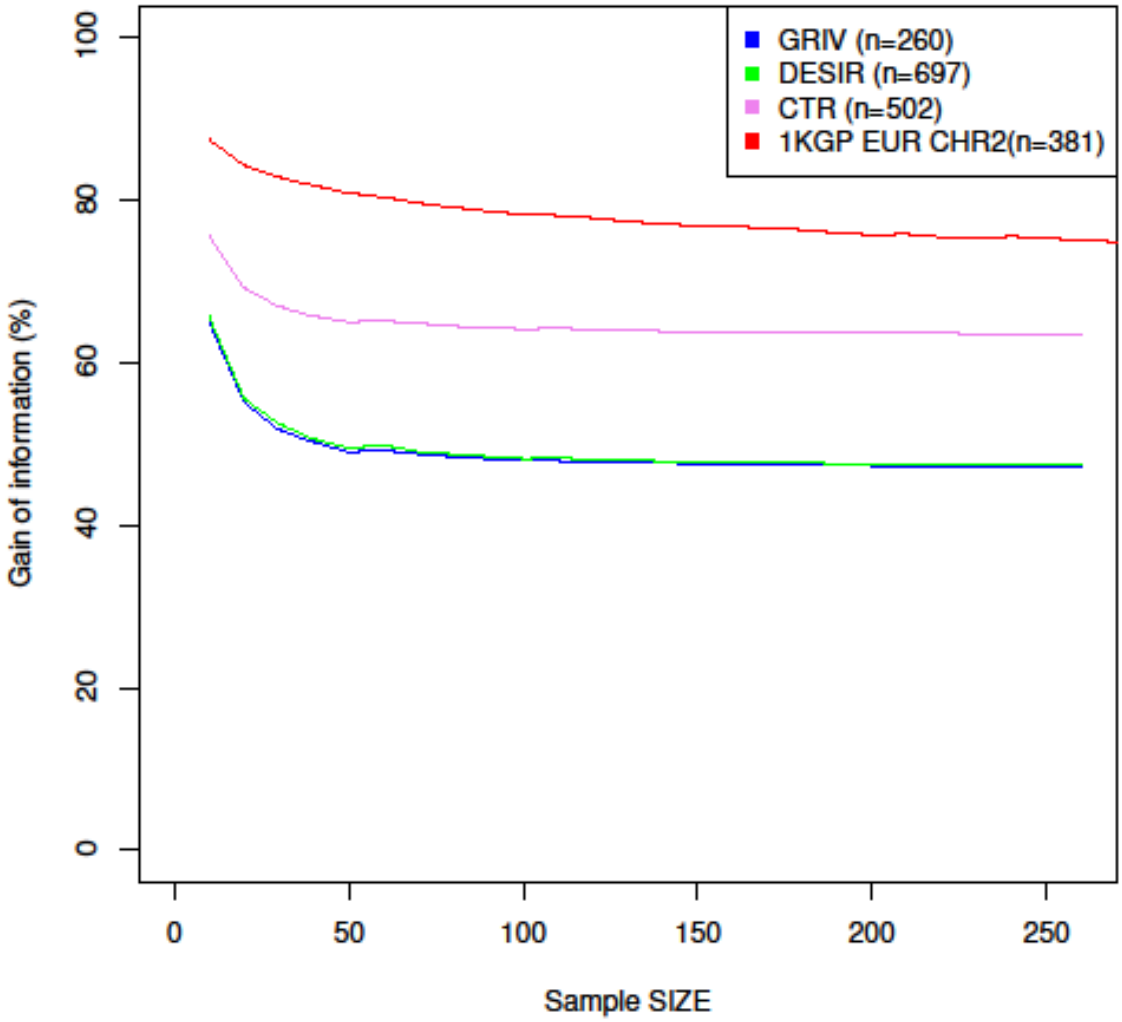

Figure 3

Référencé comme la figure 2.B dans le manuscrit de la publication.

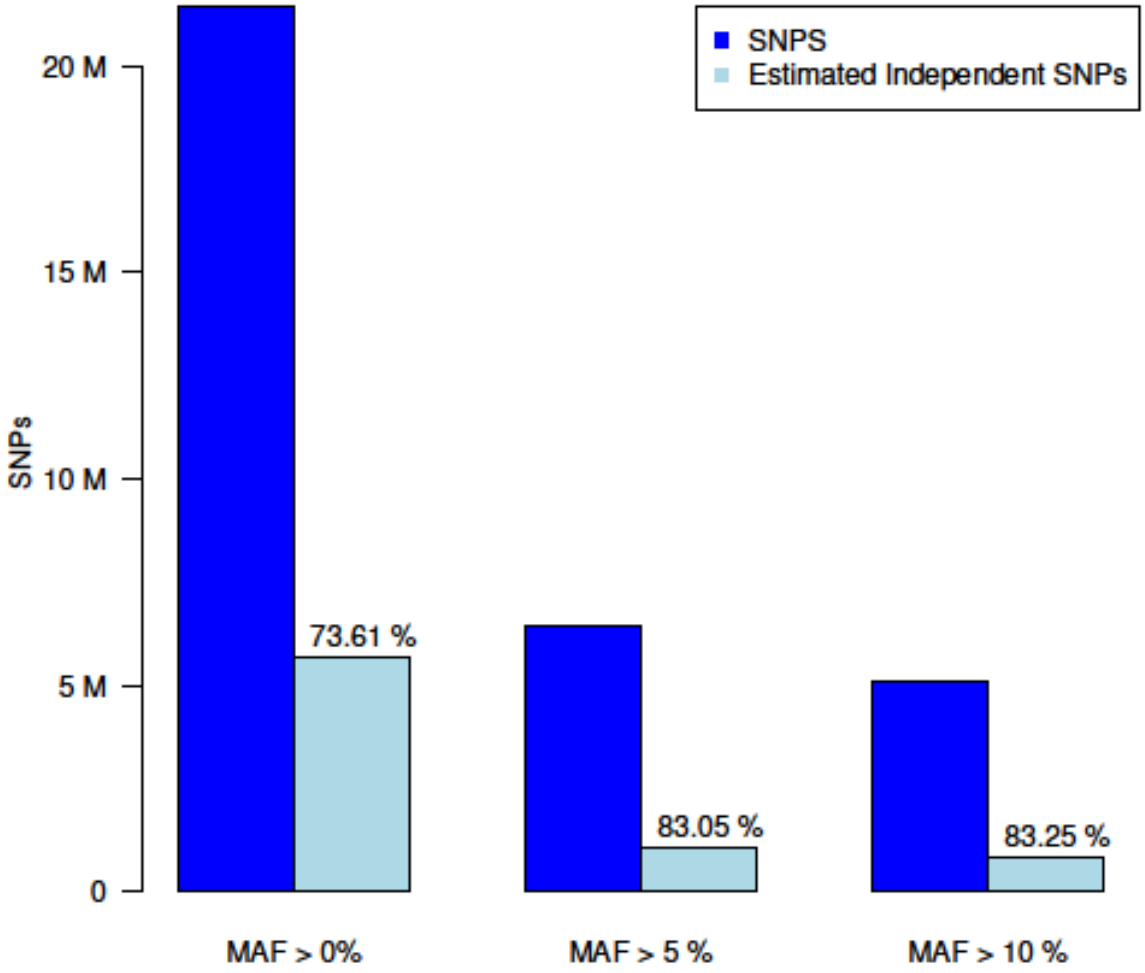

Figure 4

Référencé comme la figure 3 dans le manuscrit de la publication.

b. Résultats complémentaires

1. Tableau avec les nouveaux seuils de Bonferonni

Cohorte	SNPs	Meff Agglomeration Classique	Meff Gao	Seuil Agglomeration Classique	Seuil Gao et al
GRIV	205122	94356	147903	5,30E-07	3,38E-07
DESIR	188162	92304	146841	5,42E-07	3,41E-07
CTR	512590	192931	297779	2,59E-07	1,68E-07

Tableau 5 : Comparaison entre notre méthode et celle de gao et al. sur des données de puces de génotypage

Application de ma méthode d'agglomération classique ainsi que celle de Gao et al. sur des données issues de puces de génotypage. Les données ont été filtrées de telles sorte qu'il n'y ai aucune données manquantes ce qui explique le nombre de SNPs différents par rapport au tableau présenté dans la publication.

GENE	SNPS (M)	seuil (alpha=5%)	GAIN	Meff classique	Meff étendue	Meff Gao et al	seuil aggoleration classique	seuil aggoleration etendue	seuil Gao et al
FASLG	2	2,50E-02	60,38%	1,63	1,67	2	3,06E-02	2,99E-02	2,50E-02
IFNG	2	2,50E-02	66,98%	1,49	1,48	2	3,36E-02	3,38E-02	2,50E-02
IFNGR1	2	2,50E-02	73,08%	1,98	1,97	2	2,52E-02	2,54E-02	2,50E-02
IL1B	3	1,67E-02	63,45%	2,69	2,59	3	1,86E-02	1,93E-02	1,67E-02
IL4	3	1,67E-02	72,48%	1,67	1,66	3	3,00E-02	3,01E-02	1,67E-02
ZAP70	3	1,67E-02	80,43%	2,07	2,06	3	2,41E-02	2,43E-02	1,67E-02
IFNAR1	4	1,25E-02	75,04%	3,11	3,02	4	1,61E-02	1,66E-02	1,25E-02
LTA	4	1,25E-02	75,70%	3,13	3,13	4	1,60E-02	1,60E-02	1,25E-02
CXCR2	5	1,00E-02	78,92%	2,62	2,55	5	1,91E-02	1,96E-02	1,00E-02
GPR15	5	1,00E-02	73,08%	3,90	3,82	5	1,28E-02	1,31E-02	1,00E-02
CXCR6	6	8,33E-03	79,76%	4,11	3,93	5	1,22E-02	1,27E-02	1,00E-02
FAS	6	8,33E-03	78,85%	5,12	4,67	5	9,76E-03	1,07E-02	1,00E-02
IL10RA	6	8,33E-03	77,18%	4,27	3,53	7	1,17E-02	1,42E-02	7,14E-03
TNF	7	7,14E-03	86,84%	4,62	4,01	6	1,08E-02	1,25E-02	8,33E-03
IL6	8	6,25E-03	83,48%	4,17	3,7	6	1,20E-02	1,35E-02	8,33E-03
IL10	9	5,56E-03	88,02%	2,94	2,9	7	1,70E-02	1,72E-02	7,14E-03
IL1RN	9	5,56E-03	86,69%	3,61	3,5	5	1,39E-02	1,43E-02	1,00E-02
IL1A	10	5,00E-03	88,35%	2,76	2,7	5	1,81E-02	1,85E-02	1,00E-02
TRIM5	10	5,00E-03	75,65%	6,47	5,3	9	7,73E-03	9,43E-03	5,56E-03
IL4R	21	2,38E-03	80,99%	11,49	7,5	18	4,35E-03	6,67E-03	2,78E-03

Tableau 6 : Calcul des nouveaux seuils de Bonferroni avec les Meff sur des gènes candidats

Application de notre méthode sur des gènes candidats. Les calculs ont été faits sur les données haplotypiques du projet 1000 génomes. La méthode d'agglomération classique et

celle de Gao et al. ont considéré les données haplotypiques comme des données génomiques. Seule la méthode d'agglomération globale tire partie des haplotypes.

On remarque qu'il y a écart net entre les Meff obtenus par Gao et al et les nôtres aussi bien pour la méthode d'agglomération classique que la méthode d'agglomération étendue.

2. Illustration de la méthode étendue et comparaison avec Gao et al.

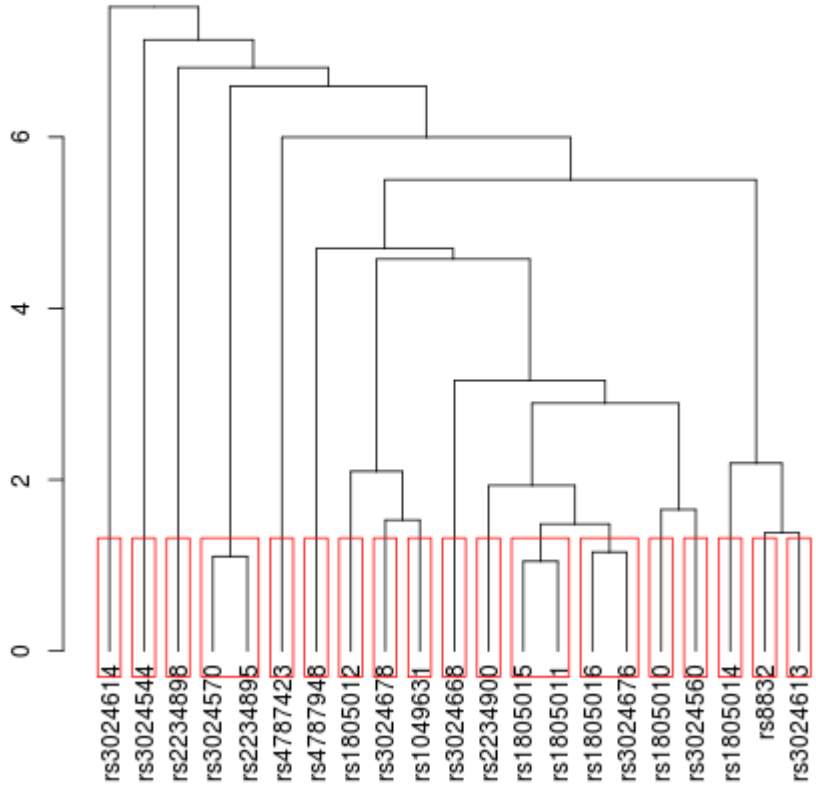

Figure 17 : Dendrogramme du gène candidat IL4R obtenue par les données intermédiaires de la méthode d'agglomération étendue et comparaison avec celle de Gao et al

En abscisse, les SNPs du gène IL4R

En ordonnée, les Meff obtenus par notre méthode d'agglomération étendue. Le nombre de Meff estimé par notre méthode est de 7,5.

Les rectangles rouges sont les Meff obtenus par la méthode de Gao et al. Ils sont supposés car la méthode ne fournit pas de données intermédiaires entre les fichiers d'entrée et les Meff calculés.

A partir des données intermédiaires de la méthode d'agglomération étendue nous pouvons reconstruire un dendrogramme des Meff intermédiaires pour mieux caractériser les motifs de déséquilibre de liaison dans le gène. Les rectangles rouges sont les Meff supposés de Gao en partant du principe que sa mesure de déséquilibre de liaison est équivalente à la nôtre et que les regroupements de SNPs sont similaires.

2. Etude génome entier sur le photo-vieillissement

a. Travail en cours de publication

Au cours de ma thèse, j'ai analysé les données génomiques obtenues par puce de génotypage sur 502 femmes issues de la cohorte SU.VI.MAX. Le but de cette étude était d'identifier des variants génétiques impliqués dans le photo-vieillissement [96]. J'ai donc réalisé l'analyse "bioinformatique" de cette étude à travers la gestion des données, les contrôles qualité ainsi que les différents tests statistiques. Cette analyse s'est montrée fructueuse puisque nous avons trouvé un signal significatif dans le SNP rs322458 dans le mode de comptage génotypique dont la p-value est $p=1.53x10^{-8}$ en prenant l'allèle A comme l'allèle de référence, avec un seuil de Bonferroni de $1.7x10^{-8}$ lors de cette analyse. En analysant sous le modèle génétique additif, on observe un coefficient de corrélation négatif (β=-0.3) entre notre compte d'allèle et notre grade. L'allèle A est donc protecteur vis-à-vis du grade de Larnier. Les coefficients de corrélation et pvalue des autres covariables sont pour l'âge (β=0.09, p-value $6.67x10^{-33}$), la classe d'IMC (β=-0.17, p-value $3x10^{-3}$), l'exposition au soleil (β=0.007, p-value 0.35), la ménopause (β=0.07, p-value=0.24), le tabac (β=0.07, p-value=0.21), le premier axe de la stratification (β=-1.56, p-value=0.07) et le second axe de la stratification (β= 0.83, p-value = 0.34).

Ce travail est en cours de publication dans la revue *Journal of Investigative Dermatology* et il vous est présenté ci-après, précédé d'un résumé en français.

Résumé de la publication

Nous avons voulu identifier les variants génétiques du photo-vieillissement. Pour cela, nous avons génotypé 502 femmes caucasiennes sur des puces de génotypage Illumina Omnione. Après les contrôles qualités, il restait près de 795 000 SNPs à analyser pour rechercher une association avec le photo-vieillissement. Des covariables comme l'âge, l'exposition au soleil, l'exposition à la cigarette, le statut hormonal et la stratification ont été prises en compte lors de l'analyse.

Après correction pour test multiple, le SNP rs322458 s'est montré significatif ($p=1.53x10^{-8}$) en mode récessif. Ce SNP est situé en amont du gène *STXBP5L* (Syntaxin binding protein 5-like), il est en déséquilibre de liaison ($R^2 > 0.8$) avec 5 SNPs (rs470647, rs612545, rs617332, rs645045 et rs1795413) localisés dans les parties introniques de ce même gène et 2 SNPs intergéniques (rs377374 et rs450614). *STXBP5L* intervient dans les processus d'exocytose et est exprimé dans la peau. Une analyse haplotypique de la région a aussi été effectuée. Il est important de noter que le SNP rs470647 ($R^2 = 0.83$ selon HapMap pour la population CEU) influe sur l'expression d'un gène voisin, *FBXO40* (F-box protein 40) [95], qui intervient dans le processus de différenciation des muscles.

A Genome-Wide Association Study in Caucasian Women Points Out for a Putative Role of the *STXBP5L* Gene in Facial Photoaging

Sigrid Le Clerc[1,11], Lieng Taing[1,11], Khaled Ezzedine[2,3], Julie Latreille[4,12], Olivier Delaneau[1,5], Toufik Labib[1], Cédric Coulonges[1], Anne Bernard[4,12], Safa Melak[1], Wassila Carpentier[6], Denis Malvy[2,7], Randa Jdid[4,12], Pilar Galan[2], Serge Hercberg[2,8], Frederique Morizot[4,12], Christiane Guinot[4,9,12], Erwin Tschachler[4,10,11,12] and Jean- François Zagury[1,11]

A genome-wide association study (GWAS) was conducted on 502 French middle-aged Caucasian women to identify genetic factors that may affect skin aging severity. A high-throughput Illumina Human Omni1-Quad beadchip was used. After single-nucleotide polymorphism (SNP) quality controls, 795,063 SNPs remained for analysis purposes. Possible stratification was first examined using the Eigenstrat method, and then the relationships between genotypes and four skin aging indicators (global photoaging, lentigines, wrinkles, and sagging) were investigated separately by linear regressions adjusted on age, smoking habits, lifetime sun exposure, hormonal status, and the two main Eigen vectors. One signal passed the Bonferroni threshold ($P=1.53\times10^{-6}$) and was significantly associated with global photoaging. It was also correlated with the wrinkling score and the sagging score. According to HapMap, this SNP, rs322458, was in linkage disequilibrium (LD) with intronic SNPs of the *STXBP5L* gene, which is expressed in the skin. In addition, it was also in LD with another SNP that increases the expression of the *FBXO40* gene in the skin. These two genes, which were not previously described in the context of aging, may constitute good candidates for the investigation of molecular mechanisms of skin photoaging.

Journal of Investigative Dermatology (2012) **0**, 000–000. doi:10.1038/jid.2012.458

INTRODUCTION

Similar to other organs, skin ages owing to passage of time. Skin aging is influenced both by inherited intrinsic factors and by extrinsic or environmental factors, such as chronic UV exposure and smoking (Malvy *et al.*, 2000; Yaar and Gilchrest, 2007). Intrinsic aging is an ineluctable process and is due to the genetically determined natural degeneration of the cell functioning and loss of extracellular matrix with age (Yaar and Gilchrest, 1990). Its clinical phenotype on the skin is mainly characterized by fine wrinkles and dry, thin, and pale skin (Fisher *et al.*, 2002; Makrantonaki and Zouboulis, 2007).

The main factor responsible for extrinsic aging of the skin is UVR. UV-induced skin aging or photoaging is defined as the premature occurrence of signs of aging on the skin, and presents with characteristic morphological changes of both the epidermal and dermal compartments (Rabe *et al.*, 2006; Yaar and Gilchrest, 2007). A number of hereditary phenotypic features influence the severity of photoaging, most notably skin color (Kligman and Kligman, 1999; Malvy *et al.*, 2000), and skin phototype (Fitzpatrick, 1988). Individuals with dark phototypes (III–IV) commonly exhibit more "hypertrophic responses" such as deep wrinkling, coarseness, and lentigines, whereas fair phototype individuals (I–II) generally show fewer wrinkles with epidermal atrophy, focal depigmentation, as well as dysplastic changes, such as actinic keratosis, nonmelanoma, and melanoma skin cancers (Rabe *et al.*, 2006; Yaar and Gilchrest, 2007; Puizina-Ivić, 2008).

Up to now, the exploration of the genes affecting skin aging has remained limited to *MC1R* gene (Elfakir *et al.*, 2010; Suppa *et al.*, 2011), or to genes involved in genetic pathologies with accelerated skin aging (Rooryck *et al.*,

[1]Équipe Génomique, Bioinformatique et Applications, Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, Paris, France; [2]UMR U557, INSERM / U1125 INRA / CNAM, University Paris 13 /Centre de Recherche en Nutrition Humaine Ile-de-France, Bobigny, France; [3]Department of Dermatology, Hôpital Saint-André, Bordeaux, France; [4]C.E.R.I.E.S., Neuilly-sur-Seine, France; [5]Department of Statistics, University of Oxford, Oxford, UK; [6]Plate-forme Post-Génomique P3S, Hôpital Pitié-Salpêtrière, Paris, France; [7]Department of Internal Medicine and Tropical Diseases, Hôpital Saint-André, Bordeaux, France; [8]Department of Public Health, Hôpital Avicenne, Bobigny, France; [9]Computer Science Laboratory, University François Rabelais, Tours, France and [10]Department of Dermatology, University of Vienna Medical School, Vienna, Austria

[11]These authors contributed equally to this work.

[12]C.E.R.I.E.S. is a research center on human skin founded by CHANEL.

Correspondence: Erwin Tschachler, Department of Dermatology, University of Vienna Medical School, Vienna, Austria. E-mail: erwin.tschachler@meduniwien.ac.at or Jean-François Zagury, Équipe Génomique, Bioinformatique et Applications, Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, 292 Rue Saint Martin, 75003 Paris, France. E-mail: zagury@cnam.fr

Abbreviations: BMI, body mass index; GWAS, genome-wide association study; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism; SU.VI.MAX, SUpplémentation en Vitamines et Minéraux AntioXydants

Received 4 July 2012; revised 28 September 2012; accepted 9 October 2012

2008; Soufir *et al.*, 2010). A candidate gene approach has previously established associations between *MC1R* gene variants, particularly loss-of-function variants, with an increased risk of severe photoaging (Elfakir *et al.*, 2010). In addition, a few studies conducted in twin cohorts have explored the associations between environmental factors, skin aging, and gene expression (Plomin *et al.*, 1994; Shekar *et al.*, 2005, 2006; Christensen *et al.*, 2009).

To unravel new genetic associations with skin aging in a systematic way, we have undertaken a genome-wide study on a well-defined sample of Caucasian women from the SU.VI.MAX (*SUpplémentation en Vitamines et Minéraux Anti-oXydants* (Antioxidant Vitamin and Mineral Supplementation)) cohort (Hercberg *et al.*, 2004). To the best of our knowledge, no genome-wide association study (GWAS) targeting skin aging in middle-aged women of European-derived ancestry has been previously reported.

RESULTS

Using the Illumina HumanOmni1-Quad BeadChips, we conducted a GWAS by testing associations between single-nucleotide polymorphisms (SNPs) and global skin photoaging on a large sample of French middle-aged women from the SU.VI.MAX cohort. After the various quality-control tests (see Materials and Methods), 795,063 genotyped SNPs were available for 502 women.

Table 1 describes the sample of women according to the severity of photoaging. We also computed the correlations between the age and the outcome variables (Table 2). We found that the correlations with age were all statistically significant ($P<0.0001$): 0.56 for the grade of photoaging, 0.61 for the score of wrinkling and the score of sagging, and 0.27 for the score of lentigines. Similarly, the correlations between the grade of photoaging and the other outcome variables were also statistically significant ($P<0.0001$): 0.78 for the score of wrinkling, 0.66 for the score of sagging, and 0.31 for the score of lentigines; the correlation between the score of wrinkling and the score of sagging reached 0.71 ($P<0.0001$; Table 2).

Our core association analysis focused on genotypic associations obtained using linear regressions, after correction for stratification and nongenetic skin aging factors. Figure 1 presents the distribution of the *P*-values obtained for each SNP along the chromosomes (Manhattan plot). One SNP located on the chromosome 3 (locus 3q13.33), rs322458, passed the Bonferroni threshold (6.28×10^{-8}) with $P=1.53\times10^{-8}$. According to HapMap, this SNP is in linkage disequilibrium (LD) with five SNPs positioned in intronic regions of the *STXBP5L* gene (rs470647, rs612545, rs617332, rs645045, and rs1795413), and with two intergenics SNPs (rs377374 and rs450614; Figure 2). A more refined analysis suggested that the effect was likely recessive. Indeed, when regrouping the individuals according to their grade of skin photoaging, the frequency of the homozygous rs3224587-AA genotype was clearly inversely proportional with photoaging severity (Figure 3): from 28% of homozygous subjects among grade 1 to 4% among grade 5. To further investigate the rs322458 SNP, we assessed its putative impact on each phenotype: lentigines, wrinkling, and sagging. No relationship was found with the lentigines score ($P=0.63$), whereas significant links were found with wrinkling and sagging scores (respectively, $P=5.6\times10^{-5}$ and $P=1.76\times10^{-3}$).

Moreover, bioinformatics databases were investigated for possible associations between SNPs and mRNA expression, regulation (splicing, polyadenylation, and miRNA), and also for putative transcription binding sites. According to Genevar (Nica *et al.*, 2011), the genotype rs470647-AA (rs470647 is in LD with the rs322458; see Figure 2) increases the expression in skin of a neighboring gene, *FBXO40* ($P=6\times10^{-4}$; Figure 4). The rs470647 SNP and *FBXO40* are at a distance of 683 kb.

To further investigate other possible associations, we also computed all the haplotypes based on two SNPs derived from both *STXBP5L* and *FBXO40* genes. Only three haplotypes were strongly associated with photoaging (Figure 4) and they implicated the rs322458 SNP. These haplotypes involved one exonic SNP and one 3′-untranslated region of the *STXBP5L* gene (respectively, rs17740066, $P=6.27\times10^{-9}$ and rs6782033, $P=3.96\times10^{-9}$), and one intronic SNP of the *FBXO40* gene (rs6775899, $P=9.52\times10^{-10}$). The rs17740066 and rs6782033 SNPs were in partial LD with rs322458 ($D'=1$); in other words, the G allele frequency of rs322458 SNP was identical with that of the haplotypes GG (rs322458–rs17740066) and GA (rs322458–rs6782033). Interestingly, the rs17740066 SNP corresponds to the Val855Ile protein variation and rs6782033 SNP corresponds to a putative binding site for a miRNA (hsa-mir-892b; Figure 4). There was no LD between the two SNPs, rs6775899 and rs322458 ($r^2=0.014$ and $D'=0.2$). However, the GA haplotype (rs322458–rs6775899) also exhibited a significant *P*-value ($P=9.52\times10^{-10}$), suggesting it might also be a haplotype of interest.

DISCUSSION

We have described here a GWAS investigating possible associations between SNPs and global skin photoaging. This research yielded an association for the rs322458 SNP connected to the *STXBP5L* gene with severity of skin photoaging, the rs322458-AA genotype being inversely linked with the severity of skin aging. This SNP was also associated with the wrinkle and sagging scores that are defined independently from the grade of photoaging, but it was not associated with the lentigines score, suggesting that: (1) its role in photoaging does not include pigmentary disorders; and (2) molecular mechanisms might be shared by sagging and wrinkling. According to the HapMap database, this SNP is also polymorphic in the Asian and African populations, and thus it would also be worth investigating these populations. As for any GWAS, additional genetic studies will be needed to affirm this association.

Another alias for *STXBP5L* is *LLGL4*, as it is homologous to the Lethal giant larvae (*Lgl*) drosophila gene (Katoh and Katoh, 2004). The protein coded by *STXBP5L* contains five WD40 repeats (or β-transducin repeats) and a C-terminal syntaxin-binding (STXB) domain. *Lgl* regulates epithelial polarity and, when mutated, may lead to tumor-like

Table 1. Description of the population according to photoaging severity

	Photoaging severity						
	Grade 1 N = 43	Grade 2 N = 86	Grade 3 N = 174	Grade 4 N = 150	Grade 5/6 [1] N = 49	Total, N = 502	P-value of test
Age (years)	50.1 ± 4.2[2]	54.1 ± 5.0	56.8 ± 5.5	60.9 ± 5.6	62.6 ± 5.2	57.6 ± 6.4	<0.0001[3]
Lifetime sun exposure (score)	5.3 ± 3.4	5.1 ± 3.5	5.2 ± 3.6	5.5 ± 3.5	5.5 ± 3.5	5.3 ± 3.5	0.84[3]
BMI classification							0.49[4]
Normal	28 (8.4)[5]	57 (17.0)	121 (36.1)	94 (28.1)	35 (10.4)	335 (66.7)	
Overweight	9 (7.4)	19 (15.6)	37 (30.3)	45 (36.9)	12 (9.8)	122 (24.3)	
Obese	6 (13.3)	10 (22.2)	16 (35.6)	11 (24.5)	2 (4.4)	45 (9.0)	
Hormonal status							<0.0001[4]
Nonmenopausal	27 (28.7)	24 (25.5)	32 (34.0)	9 (9.6)	2 (2.2)	94 (18.7)	
Menopausal with HRT	9 (3.4)	40 (15.3)	98 (37.4)	92 (35.1)	23 (8.8)	262 (52.2)	
Menopausal without HRT	7 (4.8)	22 (15.1)	44 (30.1)	49 (33.6)	24 (16.4)	146 (29.1)	
Smoking habits							0.61[4]
Never	23 (8.0)	45 (15.7)	100 (35.0)	86 (30.1)	32 (11.2)	286 (57.0)	
Former smoker	15 (9.3)	34 (21.3)	50 (31.2)	47 (29.5)	14 (8.7)	160 (31.9)	
Current smoker	5 (8.9)	7 (12.5)	24 (42.8)	17 (30.4)	3 (5.4)	56 (11.1)	
Eye color							0.21[4]
Blue/gray	14 (10.3)	18 (13.2)	50 (36.8)	36 (26.5)	18 (13.2)	136 (27.2)	
Green/hazel/brown/black	28 (7.8)	68 (18.7)	122 (33.6)	114 (31.4)	31 (8.5)	363 (72.8)	
Hair color at 20 years							0.08[4]
Blond/red	4 (3.7)	20 (18.6)	40 (37.0)	28 (25.9)	16 (14.8)	108 (21.6)	
Light and dark brown/black	38 (9.7)	66 (16.9)	132 (33.8)	122 (31.2)	33 (8.4)	391 (78.4)	
Skin color without tanning							0.78[4]
Fair	35 (9.0)	65 (16.8)	136 (35.0)	113 (29.2)	39 (10.0)	388 (77.8)	
Dark	7 (6.3)	21 (18.9)	36 (32.4)	37 (33.3)	10 (9.0)	111 (22.2)	
History of facial freckles							0.40[4]
No	25 (8.5)	55 (18.7)	104 (35.4)	87 (29.6)	23 (7.8)	294 (58.9)	
Yes	17 (8.3)	31 (15.1)	68 (33.2)	63 (30.7)	26 (12.7)	205 (41.1)	
Suntan intensity							0.71[4]
None/slight/light	23 (7.7)	49 (16.3)	101 (33.7)	96 (32.0)	31 (10.3)	300 (60.1)	
Dark/very dark	19 (9.5)	37 (18.6)	71 (35.7)	54 (27.2)	18 (9.0)	199 (39.9)	
Sunburn event frequency							0.74[4]
None/rare	28 (7.8)	61 (17.0)	123 (34.4)	113 (31.6)	33 (9.2)	358 (71.7)	
Frequent/constant	14 (9.9)	25 (17.7)	49 (34.8)	37 (26.2)	16 (11.4)	141 (28.3)	

Abbreviations: BMI, body mass index; HRT, hormonal replacement therapy.
[1]As a single woman had grade 6, she had been grouped with grade 5 individuals.
[2]Mean ±SD.
[3]Analysis of variance (ANOVA) test.
[4]The χ^2 test.
[5]Frequency and (%), because of possible missing values, the sum of the cell frequencies can be smaller than the total indicated in the top of the columns.

phenotype development. According to bioinformatics analysis (UniProt, 2011), *STXBP5L* seems to be implicated in vesicle trafficking and could have a role in exocytosis ((Katoh and Katoh, 2004; UniProt, 2011). Interestingly, *STXBP5L* has previously been associated with liver fibrosis risk in Caucasians and with chronic hepatitis C infection (Li *et al.*, 2009). *STXBP5L* is expressed in several tissues, including the skin (Safran *et al.*, 2010), and is also expressed in lung carcinoid and germ cell tumors (Katoh and Katoh, 2004).

Bioinformatics database exploration pointed out the possible role of the SNP rs322458 in the skin expression of a neighboring gene, *FBXO40*. Haplotype analysis of the

Table 2. Correlation coefficients between age and outcome variables

	Age	Score of wrinkling	Score of sagging	Score of lentigines	Grade of photoaging
Age	1	0.61	0.61	0.27	0.56
Score of wrinkling		1	0.71	0.31	0.78
Score of sagging			1	0.26	0.66
Score of lentigines				1	0.31
Grade of photoaging					1

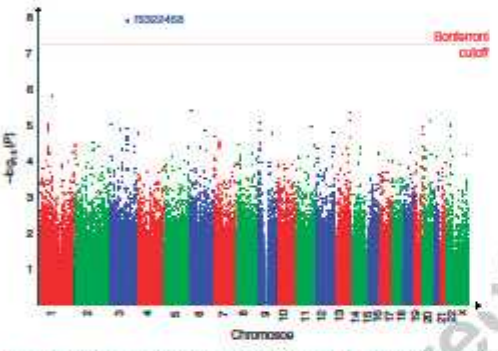

Figure 1. Manhattan plot of the association study with the photoaging score. Distribution of −log10(*P*) obtained for the associations tested between the genotypes and skin photoaging, according to the Lamier's scale, along the human chromosomes (Manhattan plot).

STXBP5L and *FBXO40* gene region also revealed positive signals ($P \sim 10^{-9}$), bringing up a second hypothesis in which rs3322458 G allele in the dominant mode, possibly in combination with other alleles, might be implicated in the phenotype.

FBXO40 encodes a protein characterized by a 40 amino-acid F-box motif. This gene is expressed specifically in the muscle (Ye *et al.*, 2007), may function as a regulator involved in the postnatal myogenesis (Ye *et al.*, 2007), and has a role in muscle hypertrophy (Shi *et al.*, 2011). F-box proteins are involved in the SCF (Skp, Cullin, F-box containing) complex, known to act as protein-ubiquitin ligases (Skowyra *et al.*, 1997), and a recent study demonstrated that the SCF–F-box40 complex prevented skeletal muscle hypertrophy by limiting the IGF1 pathway in the muscle (Shi *et al.*, 2011).

Q4

Both *STXBP5L* and *FBXO40* were not known before for any skin function. How could they affect skin aging? *FBXO40* is linked with the IGF1 pathway known for its role in inflammation, and its direct link with myogenesis could also explain its impact on wrinkling and sagging severity. Knowing that

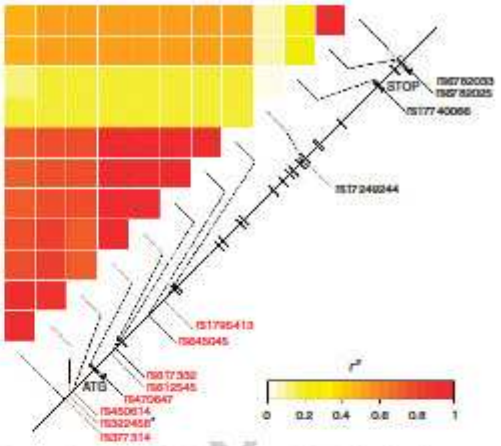

Figure 2. Genetic map of the *STXBP5L* gene. The single-nucleotide polymorphisms (SNPs) in high linkage disequilibrium (LD) with the SNP rs322458 are in red, and the exonic SNPs genotyped in the study are in black. The LD map (providing the r^2 coefficient between SNPs) is given below the genetic map. The SNP rs322458 is flagged with an asterisk (*).

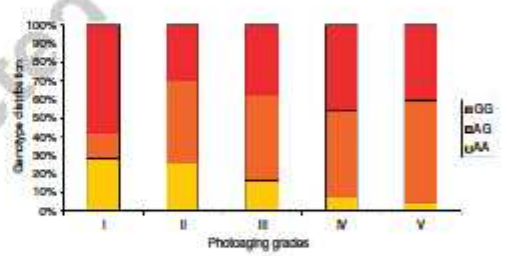

Figure 3. Distribution of the rs322458 genotypes in function of the photoaging severity.

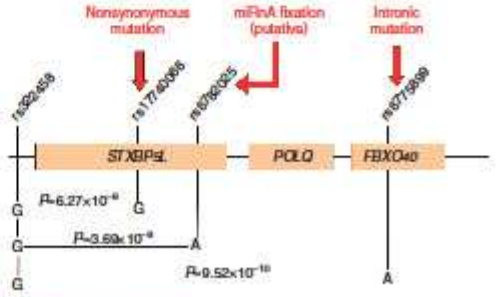

Figure 4. Haplotype analysis. All the two-single-nucleotide polymorphism (SNP) haplotypes of the region were computed using the Shapo-IT software. Associations were computed, and the figure presents the three best signals obtained, which all involve the SNP rs322458.

photoaging is intimately associated with the occurrence of dysplastic skin changes, such as actinic keratosis as well as nonmelanoma and melanoma skin cancer, it is striking to see that *STXBP5L* has been linked to cancer (Katoh and Katoh, 2004; Li *et al.*, 2009). Therefore, the search for gene polymorphisms involved in photoaging may also help to identify novel risk factors for skin carcinogenesis.

MATERIALS AND METHODS

Study design and population

A cross-sectional study was conducted to investigate skin aging in the context of the SU.VI.MAX cohort, a longitudinal cohort study, conducted in French middle-aged adults (Hercberg *et al.*, 1998). The protocol was approved by the Hospital Medicals Ethics Committee of Paris-Cochin (CCPPRB no. 706) and the "*Commission Nationale de l'Informatique et des Libertés*" (CNIL no. 334641). The study was conducted according to the Declaration of Helsinki Principles. All participants gave their written, informed consent. The SU.VI.MAX cohort included 13,017 volunteers who were representative of the French adult middle-aged population for most sociodemographic features (Hercberg *et al.*, 2004).

This study was conducted in the autumn/winter of 2002–2003. All women living in the Paris area were requested to participate in this research. Among them ($n=2,257$), 570 women, aged 44–70 years, agreed to take part in this study and provided informed consent. The participants were asked to follow specific skin care instructions; notably, application of detergents or cosmetics to the face was not authorized for at least 12 hours before the study visit. On the day of the visit, they were first asked to complete a self-administered questionnaire related to lifetime sun exposure behavior. Subsequently, three standardized, high-resolution digital images (2,008 × 3,032 pixels) of the face were taken for each participant (one frontal view of the face and one of each profile), using a Kodak DCS 760 digital camera with a 105 mm camera lens (Kodak, Paris, France). The camera was mounted on a monopod and a specifically developed chair was used to allow standardized positions of the camera with respect to the face. Lighting conditions were standardized by means of two symmetrical lamps, which provided a continuous daylight spectrum, placed at 45° to each side of the face. Finally, a blood sample was collected for genetic analysis.

Assessment of skin aging features

The facial photographs was examined for each woman by a dermatologist, and the severity of global skin photoaging was rated using a six-grade ordinal scale (Larnier *et al.*, 1994), each grade being depicted by three reference photographs that illustrate the diversity and range of pigmentation disorders, wrinkling, and sagging. In addition, the severity of 12 age-related skin features was also assessed on forehead and on cheeks using specific ordinal photographic scales (Morizot *et al.*, 2002).

Outcome variables: phenotypes analyzed

The primary outcome variable is the global photoaging grade (1–6) and the secondary outcomes variables are the three independent scores: wrinkling, sagging, and lentigines scores. On the basis of the 12 age-related skin features, the global severity of wrinkling, sagging, and solar lentigines was estimated by three scores built using principal component analysis and linear regression methods (Jobson, 1992). Then, each individual's score values were transformed to fit a range between 0 and 10.

The solar lentigines score is computed as follows: 1.25 × severity on cheeks + 1.25 × severity on forehead (with grade 0 = 0, grade 1 = 1, grade 2 = 2, grade 3 = 3, and grade >3 = 4 for each skin area). The sagging score is based on four features: 0.87 when presence of bags under the eyes + 0.78 × severity of nasolabial fold (with grade <3 = 1, grade 3 = 2, grade 4 = 3, and grade >4 = 4) + 0.93 × severity of tissue slackening + 1.07 × severity of drooping eyelids (with grade <3 = 1, grade 3 = 2, and grade >3 = 3 for the two preceding features). Finally, the wrinkling score is computed using the six remaining features: −0.64 + 0.42 × severity of wrinkles above the upper lip (with grade 0 = 0, grade 1 = 1, grade 2 = 2, grade 3 = 3, and grade 4 = 4) + 0.64 × severity of wrinkles under the eyes (with grade <3 = 0, grade 3 = 1, grade 4 = 2, and grade 5 = 3) + 0.70 × severity of fine lines on cheek (with grade 0 = 0, grade 1 = 1, and grade 2 = 2) + 0.44 × severity of furrows between eyebrows + 0.54 × severity of crow's feet + 1.06 × severity of coarse wrinkles on cheek (with grade <2 = 0, grade 2 = 1, grade 3 = 2, grade 4 = 3, and grade 5 = 4 for the three preceding features).

Covariables used for the statistical analysis: general and phenotypic data

To focus more specifically on the genetic factors affecting skin aging, several characteristics known to affect aging had to be taken into account: age (in years), body mass index (BMI; in kg m^{-2}), smoking habits (never, former, and current), and hormonal status (nonmenopausal, menopausal with hormone replacement therapy, and menopausal without hormone replacement therapy). BMI was categorized as underweight/normal (BMI <25 kg m^{-2}), overweight (25 ≤ BMI < 30 kg m^{-2}), or obese (BMI ≥30 kg m^{-2}) according to the World Health Organization (WHO) recommendations (WHO, 1995). In addition, phenotypic data such as natural hair color at the age of 20 years, eye color, skin color in winter, sunburn event frequency, suntan intensity, and history of facial freckles were also collected. Moreover, lifetime sun exposure intensity was estimated by a score based on data collected by a self-reported questionnaire. This score is a linear combination of five items weighted according to their relative contribution to the score: voluntary sun exposure, exposure of the body and the facial skin, exposure during the hottest hours of the day, intensity of self-reported lifetime sun exposure, and consideration for sunbathing. The design, validation, and description of this score have been described previously (Guinot *et al.*, 2001).

Genotyping method

The 529 women were genotyped using Illumina Infinium HumanOmni1-Quad BeadChips (Illumina, San Diego, CA) that contain 1,140,419 markers. Genomic DNA (250 ng) was whole-genome amplified, fragmented, denatured, and hybridized on prepared HumanOmni1-Quad BeadChips for a minimum of 16 hours at 48 °C. Nonspecifically hybridized fragments were removed by washing, and the remaining specific 795,063 SNPs ally hybridized DNA was fluorescently labeled by a single base extension reaction and detected using a iScan scanner (Illumina). Normalized bead-intensity data obtained for each sample were loaded into GenomeStudio software (version 1.6.3; Illumina), which converted fluorescence intensities into SNP genotypes. For the analysis, we considered only

SNPs, consequently excluding the copy-number variations that represented 91,706 markers on the HumanOmni1-Quad BeadChips. Moreover, 2,182 SNPs were removed because they were located on the Y chromosome and they could not be analyzed as the population was composed of women.

Quality control

Using the GenomeStudio software (version 1.6.3; Illumina), we analyzed the crude genotyping data, and SNPs were filtered according to the following parameters. First, nine samples with a call rate (percentage of SNPs genotyped by sample) of <95% in the Illumina clusters were removed. Second, the SNPs with a call frequency (percentage of samples genotyped by SNP) of <99% were reclustered. Third, after reclustering, samples with a call rate <98% were deleted. This method has been already used in several studies (Le Clerc et al., 2009; Limou et al., 2009, 2010). The clustering step can create SNP genotyping errors, which can be prevented by following the Illumina procedure (http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf).

This method evaluates the quality of the newly created clusters according to several criteria, which can be manually checked and corrected as necessary. In total, after all the quality control steps were carried out, 56,479 SNPs with a call frequency of <98% (2% of missing data) were excluded. This procedure ensures reliable genotyping data with little missing data. Hardy–Weinberg equilibrium analysis was performed for each SNP in each group by using an exact statistical test implemented in PLINK software (Purcell et al., 2007). Deviation from Hardy–Weinberg equilibrium in a group of patients suggests an error in genotyping. Thus, 3,866 SNPs, which were not in the Hardy–Weinberg equilibrium ($P<1.0\times10^{-3}$), were rejected in this way. We removed 191,123 SNPs with minor allele frequency <1% to avoid error of genotyping, leaving a total of 795,063 SNPs.

Identification of population stratification

To correct for possible population stratification, genotypes were analyzed using EIGENSTRAT utility of the EIGENSOFT package version 2.0 (Price et al., 2006). The two first pass with the Eigenstrat software pointed out 18 outliers, which were removed from further analyses. Then, a third pass without outliers was performed to determine the Eigen vectors. In the statistical analysis, we used the top two Eigen vectors as covariables to correct for population substructure in the association analyses (Price et al., 2006).

Statistical analysis

Of the 570 women who participated in the study, 68 were excluded from the analysis: 18 had a history of recent antiaging invasive procedures and 10 were observably non-Caucasian. In addition, one sample was removed because of insufficient DNA concentration, 12 samples were removed because the DNA was damaged, and nine samples were removed after quality control. Furthermore, 18 outliers appeared during the stratification analysis. Thus, the population investigated for our genome-wide association study was composed of 502 individuals.

The population was first described according to the severity of photoaging, using a series of analyses of variance for quantitative variables and using χ^2 tests for qualitative variables. In addition, Kendall rank correlation coefficients were calculated between age and each outcome variable, and between each pair of outcome variable (Armitage, 1971). Then, for the remaining 795,063 SNPs and 502 women, the associations between the genotypes and skin photoaging were tested. The statistical analysis was performed by a multivariate linear regression (PLINK software; Purcell et al., 2007) in the genotypic mode, taking as covariables the two first Eigenstrat principal components and the potential confounding factors (smoking habits, BMI, hormonal status, lifetime sun exposure intensity, and age). The *P*-values were adjusted by the Bonferroni correction (statistical threshold $=6.28\times10^{-8}$). Finally, for the secondary outcome variables, additional analyses were performed using the same methodology.

Haplotype inference and LD

Haplotype inference was obtained using the rapid and accurate Shape-IT algorithm (Delaneau et al., 2008, 2012). Then, for each SNP exhibiting a significant association, we looked for other SNPs in LD ($r^2>0.8$) in the HapMap population of Western European ancestry (CEU, HapMap data Release 24/phase II November 2008, on NCBI B36 assembly, dbSNP126; available at: http://www.hapmap.org) to identify the genes possibly involved with the associations. A SNP was assigned to a gene if it was located in the gene or in the 2-kb flanking regions (potential regulatory sequence); otherwise, it was considered intergenic. It is important to note that LD in HapMap population of Western European ancestry is very similar in our group of patients.

Bioinformatics exploration

To further explore the signals observed by the GWAS, we tried to look for modifications in mRNA expression levels (Genevar (Yang et al., 2010; Nica et al., 2011), Dixon (Dixon et al., 2007) databases and GHS-Express (Zeller et al., 2010) databases), splicing (NetGene2, http://www.cbs.dtu.dk/services/NetGene2/), polyadenylation regions (polyAH, http://linux1.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter and polyApred, http://www.imtech.res.in/raghava/polyapred/submission.html), transcription factor binding sites (SignalScan, http://www-bimas.cit.nih.gov/molbio/signal/, TESS, http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME, and TF Search, http://www.cbrc.jp/research/db/TFSEARCH.html, derived from TRANSFAC database), and miRNA genes or miRNA targets (miRBAse, http://www.mirbase.org/, miRTarBase, http://mirtarbase.mbc.nctu.edu.tw/, MicroCosm Targets, http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/).

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

We gratefully acknowledge the dedicated efforts of all the SU.VI.MAX volunteers, the investigators, and the staff members involved in this study, especially Dr Sandrine Bertrais, and Ms Nathalie Arnault and Mr Gwenael Menot who coordinated the data management.

REFERENCES

Armitage P (1971) *Statistical Methods in Medical Research*. Blackwell Scientific: Oxford, 504

Christensen K, Doblhammer G, Rau R *et al.* (2009) Ageing populations: the challenges ahead. *Lancet* 374:1196–208

Delaneau O, Coulonges C, Zagury JF (2008) Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 9:540

Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. *Nat Methods* 9:179–81

Dixon AL, Liang L, Moffatt MF *et al.* (2007) A genome-wide association study of global gene expression. *Nat Genet* 39:1202–7

Elkakir A, Ezzedine K, Latreille J *et al.* (2010) Functional MC1R-gene variants are associated with increased risk for severe photoaging of facial skin. *J Invest Dermatol* 130:1107–15

Fisher GJ, Kang S, Varani J *et al.* (2002) Mechanisms of photoaging and chronological skin aging. *Arch Dermatol* 138:1462–70

Fitzpatrick TB (1988) The validity and practicality of sun-reactive skin types I through VI. *Arch Dermatol* 124:869–71

Guinot C, Malvy D, Latreille J *et al.* (2001) Sun exposure behaviour of a general adult population in France. *In:* Ring J, Weidinger S, Darsow U eds *Skin and Environment - Perception and Protection.* Monduzzi editore S.p.A: Bologna, 1089–106

Hercberg S, Galan P, Preziosi P *et al.* (2004) The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Arch Intern Med* 164:2335–42

Hercberg S, Galan P, Preziosi P *et al.* (1998) Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en Vitamines et Mineraux AntiOxydants Study. *Int J Vitam Nutr Res* 68:3–20

Jobson JD (1992) *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods.* Springer Verlag: New York, 731

Katoh M, Katoh M (2004) Identification and characterization of human LLGL4 gene and mouse Llgl4 gene in silico. *Int J Oncol* 24:737–42

Kligman A, Kligman L (1999) Photoageing. In: Freeberg IM, Eisen AZ, Wolff K *et al.*, (eds) *Fitzpatrick's Dermatology in General Medicine.* McGraw-Hill: New York, 1717–23

Lamier C, Ortonne JP, Venot A *et al.* (1994) Evaluation of cutaneous photodamage using a photographic scale. *Br J Dermatol* 130:167–73

Le Clerc S, Limou S, Coulonges C *et al.* (2009) Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J Infect Dis* 200:1194–201

Li Y, Chang M, Abar O *et al.* (2009) Multiple variants in toll-like receptor 4 gene modulate risk of liver fibrosis in Caucasians with chronic hepatitis C infection. *J Hepatol* 51:750–7

Limou S, Coulonges C, Herbeck JT *et al.* (2010) Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. *J Infect Dis* 202:908–15

Limou S, Le Clerc S, Coulonges C *et al.* (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J Infect Dis* 199:419–426

Makrantonaki E, Zouboulis CC (2007) Molecular mechanisms of skin aging: state of the art. *Ann NY Acad Sci* 1119:40–50

Malvy J, Guinot C, Preziosi P *et al.* (2000) Epidemiologic determinants of skin photoaging: baseline data of the SU.VI.MAX. cohort. *J Am Acad Dermatol* 42:47–55

Morizot F, Lopez S, Guinot C *et al.* (2002) Development of photographic scales documenting features of skin ageing based on digital images. *Ann Dermatol Venereol* 129(Suppl 1 Part 2):1s402

Nica AC, Parts L, Glass D *et al.* (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 7:e1002003

Plomin R, Owen MJ, McGuffin P (1994) The genetic basis of complex human behaviors. *Science* 264:1733–9

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–9

Puizina-Ivić N (2008) Skin aging. *Acta Dermatovenerol Alp Pannonica Adriat* 17:47–54

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–75

Rabe JH, Mamelak AJ, McElgunn PJ *et al.* (2006) Photoaging: mechanisms and repair. *J Am Acad Dermatol* 55:1–19

Rooryck C, Morice-Picard F, Elcioglu NH *et al.* (2008) Molecular diagnosis of oculocutaneous albinism: new mutations in the OCA1–4 genes and practical aspects. *Pigment Cell Melanoma Res* 21:583–7

Safran M, Dalah I, Alexander J *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database (Oxford)* 2010:baq020

Shekar SN, Duffy DL, Montgomery GW *et al.* (2006) A genome scan for epidermal skin pattern in adolescent twins reveals suggestive linkage on 12p13.31. *J Invest Dermatol* 126:277–82

Shekar SN, Luciano M, Duffy DL *et al.* (2005) Genetic and environmental influences on skin pattern deterioration. *J Invest Dermatol* 125:1119–29

Shi J, Luo L, Eash J *et al.* (2011) The SCF-Fbxo40 complex induces IRS1 ubiquitination in skeletal muscle, limiting IGF1 signaling. *Dev Cell* 21:835–47

Skowyra D, Craig KL, Tyers M *et al.* (1997) F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* 91:209–19

Soufir N, Cod C, Bourillon A *et al.* (2010) A prevalent mutation with founder effect in xeroderma pigmentosum group C from north Africa. *J Invest Dermatol* 130:1537–42

Suppa M, Elliott F, Miloljevic JS *et al.* (2011) The determinants of periorbital skin ageing in participants of a melanoma case-control study in the U.K. *Br J Dermatol* 165:1011–21

UniProt (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 39:D214–9

WHO(eds) (1995) *Physical Status: The Use and Interpretation of Anthropometry.* Report of a WHO Expert Committee. WHO Technical Report Series 854 Geneva: World Health Organization

Yaar M, Gilchrest BA (1990) Cellular and molecular mechanisms of cutaneous aging. *J Dermatol Surg Oncol* 16:915–22

Yaar M, Gilchrest BA (2007) Photoageing: mechanism, prevention and therapy. *Br J Dermatol* 157:874–87

Yang TP, Beazley C, Montgomery SB *et al.* (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* 26:2474–6

Ye J, Zhang Y, Xu J *et al.* (2007) FBXO40, a gene encoding a novel muscle-specific F-box protein, is upregulated in denervation-related muscle atrophy. *Gene* 404:53–60

Zeller T, Wild P, Szymczak S *et al.* (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5:e10693

b. Etude génome entier sur le photo-vieillissement : analyses complémentaires

1. Utilisation du Meff

Après calcul du Meff selon notre méthode, nous obtenons un seuil a à 1.7×10^{-7} ce qui n'a malheureusement pas été suffisant pour déclarer de nouveaux signaux significatifs.

2. Autres phénotypes

D'autres analyses ont été faites sur les autres phénotypes disponibles à savoir un score de ride, un score de relâchement cutané et un score de lentigines. Aucun signal n'a été significatif avec la correction de Bonferroni. Nous avons donc décidé d'appliquer la correction du taux de faux positifs [70] pour ces autres analyses. Aucun signal n'est ressorti significatif lors de l'analyse du score des rides, néanmoins des signaux sont ressortis pour les autres phénotypes. Je vais présenter les résultats les plus pertinents à savoir ceux dont une recherche bibliographique ont permis de relever l'implication potentielle de gènes. Les SNPs non présentés étant des SNPs intergéniques que seulement des analyses expérimentales pourraient expliquer leur mode d'action.

Relâchement

Le SNP rs16931414 est ressorti sous le modèle génétique dominant avec une p-value de $5.84x10^{-07}$ ainsi qu'une q-value de 22.6 % (figure 16). Il se trouve en région 3' (en aval) du gène *PRRX2* (paired related homebox 2). La protéine codée par *PRRX2* est exprimée dans les fibroblastes fœtaux et dans les couches dermiques en développement avec une diminution de l'expression dans la peau chez l'adulte. Ce motif d'expression suggère un rôle dans le développement de la peau chez le fœtus.

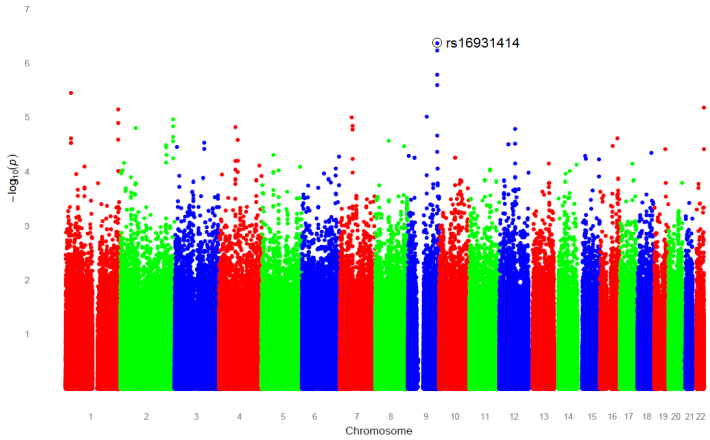

Figure 18 : Manhattan plot de l'analyse du score de relâchement

Lentigines

Le SNP rs2524069 est ressorti en mode de comptage génotypique avec une p-value de $2{,}53 \times 10^{-7}$ ainsi qu'une q-value de 8,36 % (figure 17). Le SNP est localisé dans le gène inféré *DHFRP2* (dihydrofolate reductase pseudogene 2) dont aucune fonction n'est connue. Néanmoins, le SNP est connu pour influencer des gènes du complexe majeur d'histocompatibilité à savoir *HLA-DRB1*, *HLA-DQB1*, *HLA-DQA1* et *HCG22* [59].

De plus, un autre SNP rs2853949 est aussi ressorti en mode de comptage génotypique avec une p-value de $6{,}24 \times 10^{-7}$ ainsi qu'une q-value de 8,36% (figure 17). Il est situé en aval du gène *HLA-C* appartenant lui aussi au complexe majeur d'histocompatibilité.

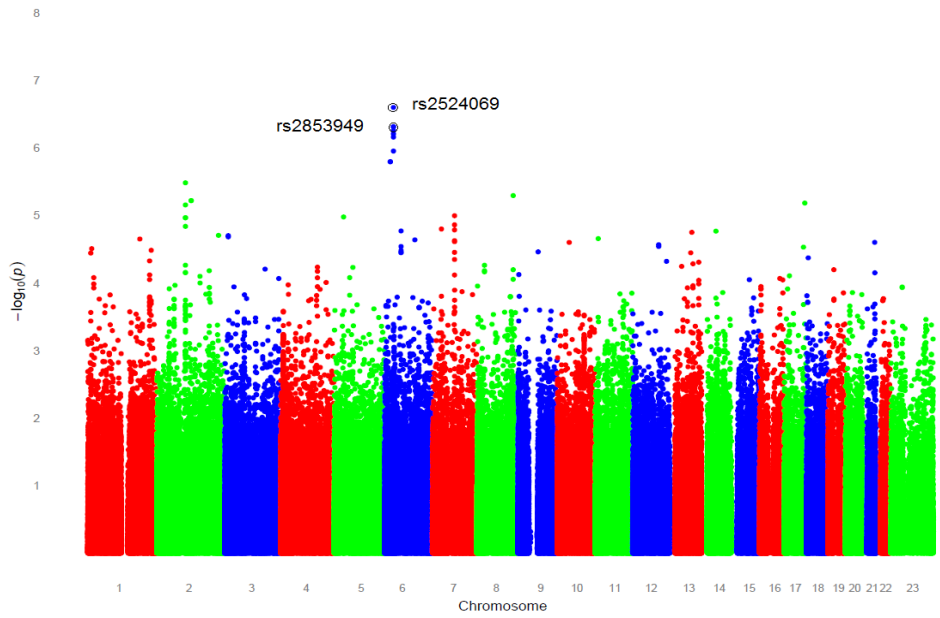

Figure 19 : Manhattan plot pour l'analyse du score de lentigines

Discussion

1. Discussion sur le logiciel Genetropy

a. Bilan de l'entropie et comparaison aux autres méthodes

Nous avons vu la possibilité de quantifier la fraction indépendante et donc aussi la fraction redondante de l'information contenue dans un jeu de données génomique. Cette fraction indépendante nous a servi comme mesure afin de déterminer le nombre de tests indépendants à considérer pour la correction des tests multiples.

Nous avons mis au point une méthode et un logiciel qui aboutissent à une mesure consistante de l'information contenue dans un jeu de données génomiques, pour une taille de population supérieure à 60 individus et à une fenêtre d'observation supérieure à 100 SNPs. Globalement, la méthode permet de gagner un facteur allant jusqu'à 80% de redondance sur des données à grande échelle, dans le cas, par exemple, des données du projet 1000 génomes. L'objectif de mon travail de bioinformaticien était en premier lieu de fournir un logiciel opérationnel utilisant l'entropie, en validant sa fiabilité informatique et la consistance des résultats obtenus. Ce logiciel est aujourd'hui opérationnel et disponible pour la communauté scientifique (http://www.griv.org/genetropy/).

Notre méthode permet d'obtenir des seuils moins stringents que celle développée par Gao et al., aussi bien sur les données de gènes candidats que sur les données de puces. Afin de voir si c'est notre correction qui est trop libérale ou celle de Gao qui est trop conservatrice, il faudrait recourir à des méthodes de permutations sur données réelles que sur données simulées, ce qui n'a pas pu être fait compte tenu du temps limité et ce n'était pas l'objectif premier de mon travail (qui était de faire un logiciel utilisable par la communauté). On peut noter que notre logiciel est moins rapide que celui de Gao (de l'ordre d'un facteur 3), mais il est plus puissant car nous avons pu traiter les données de 1000 Génomes, ce que le logiciel de Gao ne peut faire sur une machine standard.

Concernant les autres méthodes citées dans l'introduction, nous n'avons pas pu les incorporer dans cette comparaison faute de logiciels disponibles et de temps. Nous pouvons toutefois émettre certaines observations les concernant.

Ces méthodes sont basées sur une analyse par composantes principales ou plus précisément sur la décomposition spectrale. Cette méthode a un inconvénient majeur, elle n'accepte pas de données manquantes. Ceci n'est pas un handicap lorsque l'on travaille sur des données de gènes candidats mais peut être bien plus problématique lorsque l'on travaille sur des données de puces de génotypage. La moindre données manquante rends inexploitable le SNP pour l'ensemble des individus. Gao et al. utilise la méthode des K-voisins les plus proches pour reconstituer les données manquantes qui semble être une approximation suffisante pour sa méthode. A ce niveau, notre logiciel permet le calcul d'une mesure de déséquilibre de liaison avec seulement les données présentes ce qui lui permet donc d'accepter des données manquantes.

Un autre point à souligner concerne les méthodes utilisées pour calculer la matrice de corrélation plus précisément la mesure utilisée. Le déséquilibre de liaison composite permet de s'affranchir de la phase car il ne considère que les comptes alléliques de chacun des SNPs. Lors du calcul du déséquilibre de liaison composite, les fréquences sont celles des génotypes des deux SNPs et non pas celles des haplotypes (tableau 1). En tant que telle, la mesure est une approximation du déséquilibre de liaison qui n'est observable que sur des haplotypes. En comparaison à notre méthode, nous ne pouvons offrir mieux puisque la reconstitution des haplotypes par un algorithme EM donne sensiblement les mêmes résultats, c'est pourquoi nous nous efforçons d'adapter notre méthode sur des données phasées. Le fait de ne pas travailler sur les haplotypes et de ne pas prendre en compte les doubles hétérozygotes peut introduire un biais. Deux SNPs peuvent présenter les mêmes comptes génotypiques qu'ils soient en déséquilibre de liaison total ou bien en équilibre de liaison alors qu'ils devraient être considérés comme redondants dans le premier cas et pas dans le second.

De plus, le fait de considérer les SNPs uniquement par leur compte d'allèle n'est valable que dans le cadre de polymorphismes bi alléliques. Le fait de n'avoir que deux allèles impliquent que la connaissance de l'un détermine la connaissance de l'autre le compte d'allèle permet alors de renseigner le génotype dans son intégralité. Toutefois, le compte d'allèle ne permet plus de renseigner un polymorphisme multi allélique, ni de reconstruire le génotype. La mesure du déséquilibre de liaison composite n'est donc plus adaptée pour ce cas de figure. Pour ce point là, notre mesure de corrélation semble plus adaptée à des données génomiques puisqu'elle permet de prendre des polymorphismes multi alléliques en considérant chacun des allèles comme une classe et non plus comme un compte.

La méthode de Gao et al. ainsi que les autres méthodes ne semblent pas prendre en compte les effets de bord, à savoir que deux blocs consécutifs sont considérés comme étant indépendants, en d'autres termes, ces méthodes ignorent le déséquilibre de liaison entre les derniers SNPs d'un bloc et les premiers SNPs du bloc suivant. Notre méthode propose une approximation avec la moyenne de ses deux passes décalées. Ceci pourrait expliquer en partie la différence de résultats observée.

b. Perspectives : applications sur données d'haplotypes

La phase du chromosome entier, aujourd'hui possible [28], nous permettrait de nous affranchir de l'étape de phasage par paire de SNPs que nous utilisons dans notre logiciel et cette fois-ci de travailler directement sur des données pré-phasées. Dans notre travail, nous mettons en évidence que l'utilisation de données pré-phasées nous permet d'agréger non plus cette fois des paires de SNPs mais des paires de polymorphismes au sens large du terme donc entre SNPs mais aussi entre haplotypes de SNPs. Ceci nous as donc permis de caractériser un déséquilibre de liaison entre deux haplotype de 2 SNPs (fig 16 haplotype rs1805015-rs1805011 et haplotype rs1805016-rs3024676). Les résultats préliminaires utilisant des données pré-phasées ont été prometteurs. De plus, suivant les propriétés de l'entropie de Shannon, la redondance mesurée sur des haplotypes de plus de deux SNPs est supérieure ou similaire sur des haplotypes de paires de SNPs. Enfin, avec l'algorithme de Kruskal, nous pouvons alors reconstituer des blocs haplotypiques qui maximisent la redondance dans ce système en choisissant un seuil de redondance. Nous pourrions donc déterminer des blocs de redondance qui permettraient alors d'obtenir de nouveau TagSNPs en fonction d'un seuil de redondance choisi.

c. Perspectives : applications en analyse de données

L'entropie a déjà été utilisée comme critère d'analyse aussi bien pour caractériser les interactions entre SNPs dans le cadre de leur association avec un phénotype étudié [104] mais aussi comme mesure de distance entre deux distributions de fréquences alléliques ou génotypique, comme dans le cas d'un même SNP dans deux populations différentes [105]. Les données pré-phasées nous permettraient alors de caractériser des différences concernant les motifs de déséquilibre de liaison entre différentes populations. Ces différentes populations pourraient avoir des histoires migratoires différentes dans le cadre de la génétique des

populations ou bien différer sur un phénotype comme dans le cadre des études cas/contrôle. Elle permettrait de résumer avec une seule variable le déséquilibre de liaison d'une région génomique. Nous pourrions alors discriminer pour des populations à comparer et pour une région génomique des motifs de déséquilibre de liaison différents. Toutefois, le fait que l'entropie de Shanon dépende des fréquences observées entraine la possibilité d'avoir des régions génotypique avec des motifs de déséquilibre de liaison totalement différents mais en ayant la même entropie.

Enfin, au vu des nouveaux seuils a obtenus, nous restons dans le même ordre de puissance ce qui a été insuffisant aussi bien pour nos études de gènes candidats que pour nos études génome entier.

2. GWAS sur le photo-vieillissement

a. Les signaux significatifs

La GWAS sur le photo-vieillissement a pu identifier un nouveau signal sur le SNP rs322458 associé au gène *STXBP5L*, exprimé dans la peau. Il est aussi intéressant de noter que ce SNP ressort aussi dans l'analyse des phénotypes de score de rides ainsi que de score de relâchement cutané, mais pas pour le score des lentigines (respectivement $5.6x10^{-5}$, $1.76x10^{-4}$ et 0.63). Ceci nous permet de suggérer qu'il y'a un mécanisme moléculaire commun entre ces deux phénomènes.

L'utilisation de la base de données GENEVAR [106], reliant les polymorphismes alléliques et l'expression des protéines, a montré que le SNP rs322458 était aussi associé à un changement du niveau d'expression d'un gène voisin, *FBXO40*, exprimé dans les muscles. Seules des études fonctionnelles permettront de confirmer l'implication du SNP dans les mécanismes moléculaires liés au photo-vieillissement et des équipes collaboratrices ont déjà commencé à explorer l'expression du gène *STXP5L* dans différents types cellulaires cutanés. Il est aussi à souligner que le gène *STXBP5L* a aussi été retrouvé associé à des cancers tels que la fibrose du foie ou bien le carcinome pulmonaire [107]. Enfin *STXBP5L* est un homologue d'un oncogène de la drosophile *LLGL4* [107].

La GWAS a porté ses fruits puisqu'elle a permis d'identifier un signal *de novo*.

L'analyse des autres phénotypes a aussi révélé d'autres gènes potentiels tels que *PRRX2* dans les mécanismes de relâchement cutané, ainsi que le *HLA-C* pour les lentigines.

L'implication potentielle de *PRRX2* dont l'expression se fait dans les premiers stades du développement humain puis devient sous régulée au stade adulte, ainsi que son implication dans les mécanismes de cicatrisation en font une cible de choix pour d'autres études fonctionnelles. Encore une fois, son implication restait jusqu'alors inconnue dans le mécanisme du relâchement cutané. L'implication potentielle du *HLA-C* ainsi que d'autres gènes du complexe majeur d'histocompatibilité dans le score des lentigines sous-entend que le système immunitaire est impliqué dans le mécanisme d'apparition des lentigines ce qui était

chose insoupçonnée jusqu'alors. L'absence de signaux pour l'étude du score de ride peut s'expliquer par le fait que le score est en fait une synthèse de plusieurs types de rides distincts. Nous allons donc procéder à l'analyse de chacun de ces types de rides afin de préciser le phénotype.

b. Vue globale des associations identifiées

De tous ces signaux nous pouvons souligner la proximité entre nos gènes et le trait étudié. Bien que l'implication de ces gènes ne puisse être validée que par une expérimentation fonctionnelle et qu'il n'est pas impossible d'exclure un biais d'observation, ils présentent tous une proximité avec la dermatologie. En effet, ils sont exprimés dans la peau ou bien entrent dans le fonctionnement musculaire. Je tiens à mettre cela en parallèle avec les résultats obtenus au sein de Laboratoire GBA dans le cadre de l'étude GRIV montrant l'implication du chromosome 6, plus précisément la région du complexe majeur d'histocompatibilité, dans le cadre de la réponse au VIH. Une implication du système immunitaire dans le cadre d'une maladie infectieuse ayant pour cibles les cellules de ce dit système ne semble pas dénuée de sens. De manière pragmatique et superficielle, il est rassurant que les principaux résultats obtenus à l'aide des puces de génotypage aient un sens biologique lié à la problématique posée.

c. Perspectives : réplication et méta-analyse

La suite à donner classique de cette GWAS est, si possible expérimentalement, de procéder à des analyses fonctionnelle afin de confirmer nos hypothèses. Nous pouvons aussi corroborer ces résultats par le biais d'une autre GWAS étudiant les mêmes phénotypes. L'utilisation d'une autre GWAS peut se faire par réplication en se focalisant sur les p-values de SNPs d'intérêt obtenues dans d'autres GWAS comme des SNPs candidats ou bien par méta analyse en calculant une p-value combinée pour chaque SNPs communs avec une ou plusieurs autres GWAS. Cette p-value combinée pour les SNPs étudiés peut s'obtenir soit par la méthode de Fisher soit par le calcul du Z-score.

Cette réplication peut être difficile à mettre en place en fonction du phénotype étudié. En effet, plus un phénotype est défini précisément moins il est évident de le retrouver dans une autre étude. De plus, selon la construction de l'étude et plus précisément le type de puce utilisée, les SNPs communs à toutes les puces peuvent être limités et donc ne pas inclure nos

SNPs d'intérêt. Grâce à l'imputation, nous pouvons toutefois tenter de reconstruire ces SNPs manquants afin de procéder à une méta-analyse. Dans notre cas, la méta-analyse reste pour l'instant impossible puisque notre étude est la première à se focaliser sur le photo-vieillissement dans le cadre d'une GWAS et qu'il n'existe pas encore de cohorte de vérification.

3. Perspectives des GWAS

La méta-analyse ou bien la réplication ne sont pas les seules possibilités de poursuite d'une GWAS. Plusieurs méthodologies sont disponibles pour explorer de manière plus approfondie les données génomiques, que qu'on peut diviser en deux catégories : celles qui se focalisent sur des polymorphismes simples et celles qui s'attardent sur les polymorphismes multiples.

a. Polymorphismes simples

L'imputation permettant donc de reconstituer des SNPs absents des puces de génotypage à partir d'un panel de référence, permettant d'augmenter considérablement le nombre de polymorphismes analysables. De plus, le panel de référence de 1000 génomes s'est récemment étoffé d'insertions/délétions dont l'impact sur le plan biologique est plus important qu'un SNP à cause du décalage du cadre de lecture qu'il peut engendrer : protéines tronquées, sites transcriptionnels modifiées, etc.

De surcroit, les progrès concernant les puces de génotypage nous ont permis la caractérisation des CNVs. Les CNVs ont d'ailleurs déjà montré leur impact sur différentes maladies telles que la schizophrénie [108], l'autisme [109] ou bien l'obésité [110]. Toutefois, aucun consensus ne s'est constitué pour leur inférence rendant plus difficile leur analyse.

Les méthodes d'inférence reposent sur les intensités de fluorescence des SNPs présents sur les CNVs et dont l'intensité varie donc en fonction du nombre de copies portées. De nombreuses méthodes ont été développées en plus des logiciels propriétaires [111-113]. Toutefois en comparant ces méthodes avec des données réelles, on constate que les CNVs inférés sont à considérer avec précaution et qu'ils nécessitent une vérification manuelle et individuelle de ceux-ci [114].

Ces polymorphismes simples obtenus par inférence (CNVs) ou par imputation (insertions/délétions, SNPs de panels de référence) s'analysent avec la même méthodologie que les SNPs déjà présents sur les puces et ne nécessitent qu'une logistique d'inférence ou d'imputation le cas échéant.

Dans le cadre de notre étude sur le photo-vieillissement, nous avons commencé à entreprendre l'étude de l'impact de ces polymorphismes "simples". Compte-tenu de la lourdeur informatique et de la complexité des données à gérer, les résultats ne sont pas encore disponibles et ne seront donc pas intégrés dans la soutenance de mon doctorat.

b. Polymorphismes multiples

Il est aussi possible d'utiliser des méthodes exploratoires afin de prendre en compte plusieurs SNPs en même temps pour analyse. Je vais donc par la suite présenter quelques méthodes et leurs postulats biologiques sous-jacents à ces méthodes.

1. Haplotypes

Pour rappel, un haplotype est un méta *locus* de plusieurs polymorphismes localisés sur le même chromosome. L'impact d'un haplotype provient notamment du fait que les allèles des polymorphismes individuels n'aient pas d'impact sur le phénotype alors qu'une ou plusieurs combinaisons d'allèles de ces différents polymorphismes en aient un. Par exemple, si deux SNP ont des allèles impactant sur la séquence d'une protéine, chaque SNP peut ne pas avoir d'impact sur la fonctionnalité de cette protéine, par exemple dans le cas d'un récepteur donc la modification serait insuffisante pour modifier la liaison avec son ligand ; alors que la présence simultanée des deux mutations sur le récepteur vont impacter la fonctionnalité de cette protéine, pour reprendre l'exemple précédent, la liaison récepteur-ligand serait rompue ou bien altérée.

L'analyse des haplotypes soulève quelques problèmes au niveau méthodologique. Pour rappel, un haplotype constitué de n polymorphismes bi-allélique possède 2^n allèles possibles. Est ce que chacun de ces allèles possède un effet distinct sur notre phénotype ou bien seulement l'un d'entre eux ? De plus, comment définir l'haplotype que nous allons étudier ? Pourquoi inclure un polymorphisme dans l'haplotype et pas un autre ? Enfin, l'application de ces analyses en systématique sur tout le génome pose un problème sur la correction des tests multiples. Néanmoins, certaines approches ont été développées pour analyse systématique de ces haplotypes [61, 62] sur la totalité du génome.

2. Composé hétérozygote

Le composé hétérozygote est un cas particulier observable avec des haplotypes composés de deux SNPs agissant tout deux en mode récessif. Il correspond au double hétérozygote avec une mutation récessive sur chacun des chromosomes rendant ainsi les deux occurrences du gène inutilisables (figure 18). Il a fallu attendre les progrès liés au séquençage ainsi qu'à l'haplotypage pour le confirmer. Néanmoins, l'idée semble faire ses preuves comme en témoignent les quelque 680 références affiliées sur Pubmed.

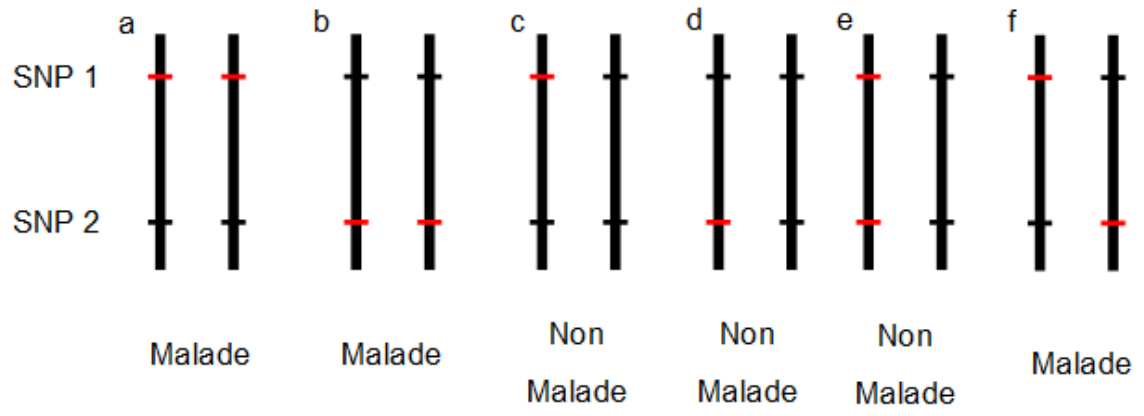

Figure 20 : Schéma descriptif d'un composé hétérozygote.

Les deux SNPs 1 et 2 présentent des polymorphismes causaux agissant en mode récessif. a) homozygote récessif SNP 1 b) homozygote récessif SNP 2 c) hétérozygote SNP 1 d) hétérozygote SNP 2 e) hétérozygote SNP 1 SNP 2 avec les allèles causaux sur le même chromosome f) hétérozygote SNP 1 et SNP 2 avec les allèles causaux sur deux chromosomes différents, appelé aussi composé hétérozygote

3. Interactions entre SNPs

Les interactions constituent l'effet de deux allèles de deux SNPs (ou plus) sur notre phénotype alors que l'effet de chacun des allèles pris indépendamment est nul ou moindre. Bien entendu le problème qui se lève assez vite correspond au problème des tests multiples puisque l'on passe au carré (grossièrement) le nombre de tests. Pourtant, de récentes études sur la maladie d'Alzheimer [115, 116] ou bien dans le cadre du diabète de type I [117] ont réussi à mettre en évidence de telles interactions ; il nous faut néanmoins souligner qu'elles se

sont restreintes à un nombre limité de SNPs, ceux liés aux processus de l'inflammation dans le cadre de l'Alzheimer et ceux du complexe majeur d'histocompatibilité pour l'étude sur le diabète.

Pour l'heure, ces approches sont donc difficilement envisageables à l'échelle du génome entier. En revanche, s'appuyer sur des a priori biologiques et cibler des voies de signalisation spécifique pourrait nous permettre de mettre en évidence de telles interactions. De nombreuses approches ont été proposées à cette effet utilisant soit l'entropie [90, 118], soit des modèles Bayesiens [119, 120], soit des arbres de décisions [121, 122]. Toutefois, la communauté [123-125] s'accorde sur le fait que le problème des tests multiples est un obstacle majeur pour la détection des interactions SNP SNP et qu'il est préférable reproduire les résultats d'interactions avec différentes méthodologies à cause de leurs spécificités propres.

4. Voies de signalisation

Bien plus qu'une simple interaction entre deux SNPs, l'analyse de voies de signalisation tire parti d'informations biologiques pré-existantes. L'idée est donc de se focaliser non pas sur l'intégralité du génome mais sur des fractions de celui-ci dont nous savons qu'elles interagissent dans une voie de signalisation. Plusieurs bases de données répertorient ces voies de signalisation comme par exemple l'encyclopédie des gènes et génomes de Kyoto (KEGG) [126] ou bien le projet BioCarta (http://www.biocarta.com). La question est donc de déterminer si une voie de signalisation est sur-représentée au niveau des résultats.

Plusieurs méthodologies ont été développées dans ce sens [127-129] présentant des différences au niveau des données d'entrée mais aussi des hypothèses testées. Deux hypothèses sont possibles :

- l'hypothèse **compétitive**, comparant les statistiques des gènes dans la voie de régulation testée à celles d'autres gènes dans le génome ;
- l'hypothèse **autonome**, comparant les statistiques des gènes dans la voie de signalisation testée à celles attendues sous hypothèse de non association.

De récentes analyses ont montré l'implication de la voie de signalisation des interleukines dans la maladie de Crohn [130, 131]. Ces voies de signalisation constituent donc l'idée que

des défaillances situées sur des gènes appartenant à la même voie sont responsables d'un même phénotype.

5. Interactions gène environnement

Au-delà des interactions directement entre nos SNPs menant à un phénotype particulier, nous pouvons aussi étudier les interactions entre nos SNPs et un facteur environnemental dont l'action combinée impacterait notre phénotype. Pour cela, il faut donc que notre étude propose à la fois des données génomiques mais aussi des données environnementales, ce qui n'est parfois pas possible. Toutefois, quelques études ont fait état de telles interactions comme dans le cas où l'utilisation d'une hormone de substitution avec un allèle du SNP rs889312 a été associée à un léger effet protecteur [132], ou bien dans le cadre d'une étude sur la dépression [133] entre le gène *5-HTTLPR* et le stress. Néanmoins, dans ces deux cas aucune p-value n'a été déclarée significative après correction des tests multiples et le nombre de polymorphismes testés est somme toute assez restreint.

c. Paradigme "variants communs, maladies communes"

Nous ressentons une certaine déception vis à vis de la GWAS, puisque nous espérions trouver davantage de signaux significatifs et aussi répliquer les signaux déjà connus et étudiés sur des études de gènes candidats. Aussi bien pour la GWAS sur le photo-vieillissement que dans le cadre du projet GRIV menée aussi au sein de mon équipe, très peu de signaux sont déclarés significatifs et les gènes dont l'implication vis à vis du phénotype était déjà connue ne sont pas ressortis. Et ce ne sont pas deux cas isolés comme en témoignent les nombreuses revues [134-138].

Bien entendu, nous ne pouvons nier les difficultés liés à la statistique notamment les problèmes de puissances liés au nombre de tests que nous effectuons. Ce problème de puissance permet entre autre d'expliquer pourquoi certains gènes dont l'association est déjà connue avec notre phénotype, ne ressortent pas dans le cadre des GWAS. Au terme de près de cinq années d'exploitation des puces de génotypage et après avoir acquis le recul nécessaire, des doutes sont soulevés vis à vis des GWAS notamment sur la capacité des polymorphismes fréquents à expliquer les maladies communes [139, 140].

L'idée donc que le variant commun est la cause d'une maladie commune est donc en train d'être remise en question [141, 142] et le dogme semble se porter sur le fait que les maladies communes soient causées par de multiples variants aux fréquences faibles avec de très fort effets proches d'un effet mendelien [143]. Cependant, ces SNPs de fréquences faibles (inférieure à 1%) ne sont pas caractérisables par les puces puisqu'ils n'en font pas partie. On comprend alors que les puces de génotypage ne peuvent apporter de réponses convenables à ce nouveau paradigme. D'une part, ces polymorphismes rares auront quelques difficultés à être caractérisés, d'autre part, il faudrait que ces mêmes polymorphismes soient déjà référencés sur ces puces. Dès lors, il nous faut nous tourner vers de nouvelles technologies permettant de caractériser ces SNPs de fréquence faible.

d. Avancées technologiques

Les progrès dans les technologies de séquençage ont permis de déterminer le génotype de SNPs fréquences rares, et d'indels sur l'intégralité du génome. De plus, cette avancée technologique s'est accompagnée d'une baisse des coûts et constitue ce qui est appelé la prochaine génération du séquençage, "Next Generation Sequencing" en anglais et abrégé en NGS. Pour l'heure, elle se divise principalement en deux "courants" qui diffèrent au niveau de la cible du séquençage.

1. Séquençage intégral

Le séquençage peut se faire sur l'intégralité du génome, on parle alors séquençage intégral du génome ou "whole-genome sequencing" en anglais. Quelques études ont appliqué ce séquençage intégral afin d'identifier les variants causaux comme dans le cadre d'une analyse de liaison chez une famille atteinte du syndrome de Miller [144] mettant en évidence un gène (*DHODH*) correspondant aux contraintes de fréquences attendues ainsi qu'un motif de ségrégation compatible avec celui du trait au sein de la famille. Une autre étude de liaison a mis en évidence un gène (*ABCG5*) susceptibles d'entrer dans le cadre de l'hyper-cholestérolémie sévère [145]. Enfin, une troisième étude de liaison a mis en évidence une mutation non synonyme dans le gène *SH3TC2* dans la neuropathie Charcot-Marie-Tooth [146].

Il est à noter que les études citées sont des études de liaison qui permettent de cibler de manière plus efficace les gènes se ségrégeant avec la maladie, et permettant aussi de travailler

avec un nombre bien plus restreint de patients. Le séquençage génome entier reste encore peu efficace sur des maladies polygéniques à cause du nombre impressionnant de variants dans le génome humain et du manque de puissance pour les analyses.

2. Séquençage de l'exome

Le séquençage de l'exome se focalise sur une partie restreinte du génome humain, sa partie fonctionnelle. Elle est donc moins onéreuse et propose une couverture (au sens de la fiabilité des séquences) bien plus importante que son homologue génome entier [147, 148]. De surcroit, le fait de se restreindre à une partie fonctionnelle et annotée du génome permet une interprétation plus simple des résultats obtenus. Ce sont les raisons pour lesquelles le séquençage de l'exome semble constituer le successeur des GWAS et jouit d'une très grande popularité.

Bien que relativement récentes, les études sur l'exome ont eu quelques succès tels que l'identification de nouvelles cibles *de novo* dans le cadre de la schizophrénie [149-151] ou bien de l'autisme [149, 152-155]. Ces succès sont principalement des études de liaisons mais le séquençage de l'exome connait aussi quelques succès dans le cadre des études d'association notamment dans le cancer [156, 157].

Néanmoins, ces nouvelles technologies sont aussi accompagnées de leur défis sur différents plans tels que la bioinformatique, l'informatique, la statistique et enfin le coût :

- la bioinformatique concernant le séquençage à haut débit doit prendre en compte la lecture de reads (fragments de génomes amplifiées et séquencés), l'alignement de ce ceux-ci sur le génome et enfin l'annotation de ces reads avant de pouvoir obtenir les génotypes ;

- les ressources informatiques nécessaires pour prendre en charge les données de séquençage que ce soit en génome entier ou bien au niveau de l'exome n'ont plus aucun rapport avec les données de puces de génotypage. Là où le giga octet était suffisant pour stocker les données d'une cohorte entière, il est insuffisant pour stocker les informations d'un seul individu. De plus les besoins nécessaires aux nouveaux programmes bioinformatiques ont aussi augmenté ;

- Au niveau des statistiques, en changeant de dogme, nous ne nous focalisons plus sur des SNPs individuellement mais sur les unités fonctionnelles que sont les gènes ce qui nécessite le développement (ou bien l'application) de tests plus adaptés ;

- Enfin d'un point de vue financier, bien que les coûts du séquençage aient baissé (de 100 millions à 10 milliers de dollars en moyenne par individu en 10 ans), ils restent bien plus chers que pour une puce de génotypage (de l'ordre d'un pour dix) sans compter les ressources informatiques.

Toutefois, le séquençage et les variants rares ne sont pas la panacée de la génomique. De récentes études remettent en cause l'hypothèse selon laquelle les maladies fréquentes seraient liées à des variants [158, 159] et penchent plus vers un modèle mixte comprenant à la fois des variants rares mais aussi des variants fréquents [138, 158].

e. Autres technologies

Au-delà de la génomique au sens strict du terme, il ne faut pas oublier que le génome n'est pas simplement limité à une succession de nucléotides. Il comprend aussi d'autres variations au niveau de sa structure, notamment la méthylation des cytosines ou bien des histones que l'on nomme épigénétique. Encore peu connu à l'heure actuelle, l'épigénétique est transmissible de manière héréditaire mais peut aussi apparaitre spontanément aux conditions environnementales.

De plus, bien au-delà de la génomique, il faut aussi considérer la transcriptomique ainsi que la protéomique qui sont respectivement l'étude à haut débit des ARNs et celles des protéines produites. De surcroit, le transcriptome partage lui aussi les progrès technologiques permettant désormais une couverture exhaustive de tout les ARNs, là où il s'appuyait sur des puces auparavant. Enfin, ces deux méthodes présentent deux intérêts par rapport à la génomique :

- elles sont localisées, à savoir que selon le tissu étudié, les résultats ne seront pas les mêmes contrairement au génome pour lequel ils restent identiques ;

- elles sont temporelles, à savoir que l'observation est dépendante du moment où elle a été réalisée.

L'étude des produits du génome permet donc un dosage plus fin des produits et ainsi donc d'avoir une meilleure idée de l'impact qu'ils peuvent avoir sur le phénotype. Toutefois, elles nécessitent plus de prélèvements que pour la génomique.

Quoi qu'il en soit, malgré les relatives déceptions qui lui sont liées, la GWAS reste une étape incontournable de la génomique. Peu onéreuse, avec une méthodologie solide elle constitue le premier coup de sonde permettant de déterminer une région d'intérêt. La GWAS a prouvé de par ses résultats l'efficacité de l'interdisciplinarité. La convergence de plusieurs disciplines telles que la génétique, l'informatique et bien entendu la statistique constitue désormais un progrès irréversible promis à un avenir des plus engageants.

Conclusion

Depuis mon arrivée au sein de l'équipe GBA, il y a de cela près de 5 ans, j'ai pu assister à l'essor des analyses de données génomiques à haut débit. J'ai pu suivre les progrès constants concernant le nombre de polymorphismes caractérisables passant d'une centaine de milliers aux millions d'aujourd'hui. La combinaison des besoins informatiques, des méthodologies statistiques adéquates, et des problématiques biologiques sous-jacentes, font de la bioinformatique une discipline récente, encore à ses premiers balbutiements.

Les travaux exposés dans cette thèse correspondent à une maîtrise de méthodes ainsi qu'à un travail de méthodologie exploratoire. Ils sont tous les deux les reflets de la discipline, à savoir pour la GWAS une application d'un protocole établi et éprouvé. Concernant le logiciel Genetropy nous proposons une approche novatrice et exploratoire face aux problématiques des analyses de demain.

L'application des méthodologies établies m'ont permis l'identification d'un SNP rs322458 associé au photo-vieillissement, situé près du gène *STXBP5L* et aussi influençant l'expression du gène *FBXO40*. Ces résultats ont permis de formuler deux hypothèses concernant le photo-vieillissement : en premier, l'indépendance avec le phénotype du score des lentigines avec le photo-vieillissement et en second, que des mécanismes moléculaires communs seraient partagés entre les mécanismes de formation des rides et ceux liés au relâchement cutané.

La méthodologie développée pour le logiciel Genetropy a permis de quantifier précisément la redondance dans un jeu de données génomiques. Bien qu'à l'heure actuelle son utilisation pour les problèmes de tests multiples reste sans impact dans la ré-analyse de nos résultats, nous restons confiants quant à son apport pour l'analyse des données génomiques. Ses perspectives d'utilisation nous incitent à poursuivre dans cette voie en l'adaptant pour traiter directement des données phasées car nous pensons que cela permettra de répondre aux problématiques de demain concernant le déséquilibre de liaison.

Pour conclure, cette thèse m'aura permis d'élargir mes connaissances sur différentes disciplines qui, il y a encore quelques années, restaient encore cloisonnées entre elles. Mon travail au cœur de cette synergie a été plaisant, formateur et particulièrement enrichissant. Il m'a donné l'impression d'être en plein milieu d'une voie rapide où les découvertes s'enchaînent aussi bien au niveau méthodologique que technologique et où ce qui était encore impensable quelques années auparavant est en passe de devenir le standard. Tout ceci pour conclure sur le fait que la bioinformatique est une discipline en pleine expansion et promise à un avenir des plus radieux.

Bibliographie

[1] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al.. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409, p.860-921.

[2] Sybert VP, McCauley E. Turner's syndrome. *New England Journal of Medecine*, 2004, 351, p.1227-1238.

[3] Klinefelter HF. Klinefelter's syndrome: historical background and development. *Southern Medical Journal*, 1986, 79, p.1089-1093.

[4] Sullivan BA, Schwartz S, Willard HF. Centromeres of human chromosomes. *Environmental and Molecular Mutagenesis*, 1996, 28, p.182-191.

[5] Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H et al.. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (ctg) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell*, 1992, 68, p.799-808.

[6] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 2010, 467, p.1061-1073.

[7] Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H et al.. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 2008, 451, p.998-1003.

[8] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD et al.. Global variation in copy number in the human genome. *Nature*, 2006, 444, p.444-454.

[9] The International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL et al.. A second generation human haplotype map of over 3.1 million snps. *Nature*, 2007, 449, p.851-861.

[10] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 2005, 437, p.1299-1320.

[11] The International HapMap Consortium. The international hapmap project. *Nature*, 2003, 426, p.789-796.

[12] Hardy GH. Mendelian proportions in a mixed population. *Science*, 1908, 28, p.49-50.

[13] Weinberg W. Über den nachweis der vererbung beim menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, 1908, 64, p.368-382.

[14] Behar DM, Harmant C, Manry J, van Oven M, Haak W, Martinez-Cruz B et al.. The basque paradigm: genetic evidence of a maternal continuity in the franco-cantabrian region since pre-neolithic times. *American Journal of Human Genetics*, 2012, 90, p.486-493.

[15] Culleton R, Coban C, Zeyrek FY, Cravo P, Kaneko A, Randrianarivelojosia M et al.. The origins of african plasmodium vivax; insights from mitochondrial genome sequencing. *PLoS One*, 2011, 6, p.e29137.

[16] Keijser S, Kurreeman FAS, de Keizer RJW, Dogterom-Ballering H, van der Lelij A, Jager MJ et al.. Il-10 promotor haplotypes associated with susceptibility to and severity of bacterial corneal ulcers. *Experimental Eye Research*, 2009, 88, p.1124-1128.

[17] Assaf A, Hoang TV, Faik I, Aebischer T, Kremsner PG, Kun JFJ et al.. Genetic evidence of functional ficolin-2 haplotype as susceptibility factor in cutaneous leishmaniasis. *PLoS One*, 2012, 7, p.e34113.

[18] Robbins RB. Some applications of mathematics to breeding problems iii. *Genetics*, 1918, 3, p.375-389.

[19] Lewontin RC. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 1964, 49, p.49-67.

[20] Hill W, Robertson A. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 1968, 38, p.226-231.

[21] Jennings HS. The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. *Genetics*, 1917, 2, p.97-154.

[22] Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature Review Genetics*, 2008, 9, p.477-485.

[23] Mueller JC. Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics*, 2004, 5, p.355-364.

[24] Clark AG. Inference of haplotypes from pcr-amplified samples of diploid populations. *Molecular Biology and Evolution*, 1990, 7, p.111-122.

[25] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 1977, 39, p.1-38.

[26] Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 1995, 12, p.921-927.

[27] Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 2001, 68, p.978-989.

[28] Delaneau O, Marchini J, Zagury J. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 2012, 9, p.179-181.

[29] Delaneau O, Coulonges C, Zagury J. Shape-it: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 2008, 9, p.540.

[30] Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nature Review Genetics*, 2011, 12, p.703-714.

[31] Gao G, Allison DB, Hoeschele I. Haplotyping methods for pedigrees. *Human Heredity*, 2009, 67, p.248-266.

[32] Niu T. Algorithms for inferring haplotypes. *Genetic Epidemiology*, 2004, 27, p.334-347.

[33] Salem RM, Wessel J, Schork NJ. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Human Genomics*, 2005, 2, p.39-66.

[34] Browning BL, Browning SR. Haplotypic analysis of wellcome trust case control consortium data. *Human Genetics*, 2008, 123, p.273-280.

[35] Lejeune J, Turpin R, Gautier M. [mongolism; a chromosomal disease (trisomy)]. *Bulletin de l'Académie Nationale de Médecine*, 1959, 143, p.256-265.

[36] Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M et al.. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 1989, 245, p.1059-1065.

[37] Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z et al.. Identification of the cystic fibrosis gene: cloning and characterization of complementary dna. *Science*, 1989, 245, p.1066-1073.

[38] Bertram L, Tanzi RE. The genetic epidemiology of neurodegenerative disease. *Journal of Clinical Investigation*, 2005, 115, p.1449-1457.

[39] Nakamura S. [huntington's disease--advances in gene mapping]. *Nihon Rinsho*, 1993, 51, p.2481-2487.

[40] Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE et al.. A polymorphic dna marker genetically linked to huntington's disease. *Nature*, 1983, 306, p.234-238.

[41] Coon H, Matsunami N, Stevens J, Miller J, Pingree C, Camp NJ et al.. Evidence for linkage on chromosome 3q25-27 in a large autism extended pedigree. *Human Heredity*, 2005, 60, p.220-226.

[42] Autism Genome Project Consortium. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genetics*, 2007, 39, p.319-328.

[43] Karlsson R, Graae L, Lekman M, Wang D, Favis R, Axelsson T et al.. Magi1 copy number variation in bipolar affective disorder and schizophrenia. *Biological Psychiatry*, 2012, en presse.

[44] Melhem N, Middleton F, McFadden K, Klei L, Faraone SV, Vinogradov S et al.. Copy number variants for schizophrenia and related psychotic disorders in oceanic palau: risk and transmission in extended pedigrees. *Biological Psychiatry*, 2011, 70, p.1115-1121.

[45] Namiki T, Tanemura A, Valencia JC, Coelho SG, Passeron T, Kawaguchi M et al.. Amp kinase-related kinase nuak2 affects tumor growth, migration, and clinical outcome of human melanoma. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108, p.6597-6602.

[46] Wu X, Ye Y, Rosell R, Amos CI, Stewart DJ, Hildebrandt MAT et al.. Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. *Journal of the National Cancer Institute*, 2011, 103, p.817-825.

[47] Xun WW, Brennan P, Tjonneland A, Vogel U, Overvad K, Kaaks R et al.. Single-nucleotide polymorphisms (5p15.33, 15q25.1, 6p22.1, 6q27 and 7p15.3) and lung cancer survival in the european prospective investigation into cancer and nutrition (epic). *Mutagenesis*, 2011, 26, p.657-666.

[48] Morgan TM, House JA, Cresci S, Jones P, Allayee H, Hazen SL et al.. Investigation of 95 variants identified in a genome-wide study for association with mortality after acute coronary syndrome. *BMC Medical Genetics*, 2011, 12, p.127.

[49] Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A et al.. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet*, 2010, 376, p.1393-1400.

[50] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS et al.. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106, p.9362-9367.

[51] Le Clerc S, Coulonges C, Delaneau O, Van Manen D, Herbeck JT, Limou S et al.. Screening low-frequency snps from genome-wide association study reveals a new risk allele for progression to aids. *Journal of Acquired Immune Deficiency Syndromes*, 2011, 56, p.279-284.

[52] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A et al.. Genes mirror geography within europe. *Nature*, 2008, 456, p.98-101.

[53] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 2003, 164, p.1567-1587.

[54] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 2000, 155, p.945-959.

[55] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 2006, 38, p.904-909.

[56] Bacanu S, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. *Genetic Epidemiology*, 2002, 22, p.78-93.

[57] Devlin B, Roeder K. Genomic control for association studies. *Biometrics*, 1999, 55, p.997-1004.

[58] Yu GX, Snyder EE, Boyle SM, Crasta OR, Czar M, Mane SP et al.. A versatile computational pipeline for bacterial genome annotation improvement and comparative analysis, with brucella as a use case. *Nucleic Acids Research*, 2007, 35, p.3953-3962.

[59] Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC et al.. A genome-wide association study of global gene expression. *Nature Genetics*, 2007, 39, p.1202-1207.

[60] Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K et al.. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proceedings of the National Academy of Sciences of the United States of America*, 2000, 97, p.10483-10488.

[61] Kenny EE, Gusev A, Riegel K, Lütjohann D, Lowe JK, Salit J et al.. Systematic haplotype analysis resolves a complex plasma plant sterol locus on the micronesian island of kosrae. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106, p.13886-13891.

[62] Trégouët D, König IR, Erdmann J, Munteanu A, Braund PS, Hall AS et al.. Genome-wide haplotype association study identifies the slc22a3-lpal2-lpa gene cluster as a risk locus for coronary artery disease. *Nature Genetics*, 2009, 41, p.283-285.

[63] Rosenberg N, Murata M, Ikeda Y, Opare-Sem O, Zivelin A, Geffen E et al.. The frequent 5,10-methylenetetrahydrofolate reductase c677t polymorphism is associated with a common haplotype in whites, japanese, and africans. *American Journal of Human Genetics*, 2002, 70, p.758-762.

[64] Kidd JR, Friedlaender F, Pakstis AJ, Furtado M, Fang R, Wang X et al.. Single nucleotide polymorphisms and haplotypes in native american populations. *American Journal of Physical Anthropology*, 2011, 146, p.495-502.

[65] Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 2010, 34, p.816-834.

[66] Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 2007, 39, p.906-913.

[67] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 2009, 5, p.e1000529.

[68] Fisher RA. Statistical methods for research workers. Oliver and Boyd (Ed.). Edinburgh, 1925.

[69] Stouffer S, Suchman E, DeVinney L, Star S, Williams RJ. Adjustment during army life. Oxford E (Ed.). Princeton Univ. Press., 1949.

[70] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 1995, 57 (1), p.289–300.

[71] Forner K, Lamarine M, Guedj M, Dauvillier J, Wojcik J. Universal false discovery rate estimation methodology for genome-wide association studies. *Human Heredity*, 2008, 65, p.183-194.

[72] Guedj M, Robelin D, Hoebeke M, Lamarine M, Wojcik J, Nuel G. Detecting local high-scoring segments: a first-stage approach for genome-wide association studies. *Statistical Applications in Genetics and Molecular Biology*, 2006, 5, p.Article22.

[73] Dalmasso C, Bar-Hen A, Broët P. A constrained polynomial regression procedure for estimating the local false discovery rate. *BMC Bioinformatics*, 2007, 8, p.229.

[74] Kimmel G, Shamir R. A fast method for computing high-significance disease association in large population-based studies. *American Journal of Human Genetics*, 2006, 79, p.481-492.

[75] Browning BL. Presto: rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics*, 2008, 9, p.309.

[76] Conneely KN, Boehnke M. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *American Journal of Human Genetics*, 2007, 81, p.1158-1168.

[77] Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, 2008, 9, p.516.

[78] Weir BS, Hill WG, Cardon LR. Allelic association patterns for a dense snp map. *Genetic Epidemiology*, 2004, 27, p.442-450.

[79] Gao X. Multiple testing corrections for imputed snps. *Genetic Epidemiology*, 2011, 35, p.154-158.

[80] Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, 2010, 34, p.100-105.

[81] Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 2008, 32, p.361-369.

[82] Cheverud JM. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity (Edinb)*, 2001, 87, p.52-58.

[83] Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)*, 2005, 95, p.221-227.

[84] Galwey NW. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 2009, 33, p.559-568.

[85] Dudbridge F, Koeleman BPC. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American Journal of Human Genetics*, 2004, 75, p.424-435.

[86] Pe'er I, de Bakker PIW, Maller J, Yelensky R, Altshuler D, Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, 2006, 38, p.663-667.

[87] Shannon CE. The mathematical theory of communication. 1963. *M.D. computing : Computers in Medical Practice*, 1997, 14, p.306-317.

[88] Liang Y, Kelemen A. Sequential support vector regression with embedded entropy for snp selection and disease classification. *Statistical Analysis and Data Mining*, 2011, 4, p.301-312.

[89] Ruiz-Marín M, Matilla-García M, Cordoba JAG, Susillo-González JL, Romo-Astorga A, González-Pérez A et al.. An entropy test for single-locus genetic association analysis. *BMC Genetics*, 2010, 11, p.19.

[90] Miller DJ, Zhang Y, Yu G, Liu Y, Chen L, Langefeld CD et al.. An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics*, 2009, 25, p.2478-2485.

[91] Nothnagel M, Fürst R, Rohde K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Human Heredity*, 2002, 54, p.186-198.

[92] Zhang L, Liu J, Deng H. A multilocus linkage disequilibrium measure based on mutual information theory and its applications. *Genetica*, 2009, 137, p.355-364.

[93] Yaar M, Gilchrest BA. Ageing and photoageing of keratinocytes and melanocytes. *Clinical and Experimental Dermatology*, 2001, 26, p.583-591.

[94] Yaar M, Gilchrest BA. Photoageing: mechanism, prevention and therapy. *The British Journal of Dermatology*, 2007, 157, p.874-887.

[95] Malvy JM, Guinot C, Preziosi P, Vaillant L, Tenenhaus M, Galan P et al.. Epidemiologic determinants of skin photoaging: baseline data of the su.vi.max. cohort. *Journal of the American Academy of Dermatology*, 2000, 42, p.47-55.

[96] Larnier C, Ortonne JP, Venot A, Faivre B, Béani JC, Thomas P et al.. Evaluation of cutaneous photodamage using a photographic scale. *The British Journal of Dermatology*, 1994, 130, p.167-173.

[97] Hercberg S, Galan P, Preziosi P, Roussel AM, Arnaud J, Richard MJ et al.. Background and rationale behind the SU.VI.MAX study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. supplementation en vitamines et minéraux antioxydants study. *International Journal for Vitamin and Nutrition Research*, 1998, 68, p.3-20.

[98] Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D et al.. The su.vi.max study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Archives of Internal Medicine*, 2004, 164, p.2335-2342.

[99] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al.. Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 2007, 81, p.559-575.

[100] Hodgson C. Senile lentigo. *Archives of Dermatology*, 1963, 87, p.197-207.

[101] Cario-Andre M, Lepreux S, Pain C, Nizard C, Noblesse E, Taïeb A. Perilesional vs. lesional skin changes in senile lentigo. *Journal of Cutaneous Pathology*, 2004, 31, p.441-447.

[102] Jobson JD. Applied multivariate data analysis. Springer (Ed.). , 1992.

[103] Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, 2010, 11, p.499-511.

[104] Cui Y, Kang G, Sun K, Qian M, Romero R, Fu W. Gene-centric genomewide association study via entropy. *Genetics*, 2008, 179, p.637-650.

[105] Zhao J, Boerwinkle E, Xiong M. An entropy-based statistic for genomewide association studies. *American Journal of Human Genetics*, 2005, 77, p.27-40.

[106] Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M et al.. The architecture of gene regulatory variation across multiple human tissues: the muther study. *PLoS Genetics*, 2011, 7, p.e1002003.

[107] Katoh M, Katoh M. Identification and characterization of human llgl4 gene and mouse llgl4 gene in silico. *International Journal of Oncology*, 2004, 24, p.737-742.

[108] Cook EHJ, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. *Nature*, 2008, 455, p.919-923.

[109] Kakinuma H, Sato H. Copy-number variations associated with autism spectrum disorder. *Pharmacogenomics*, 2008, 9, p.1143-1154.

[110] Fernando MMA, de Smith AJ, Coin L, Morris DL, Froguel P, Mangion J et al.. Investigation of the hin200 locus in uk sle families identifies novel copy number variants. *Annals of Human Genetics*, 2011, 75, p.383-397.

[111] Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide snp genotyping. *Nature Genetics*, 2008, 40, p.1199-1203.

[112] Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA et al.. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Research*, 2007, 17, p.1665-1674.

[113] Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S. Sparse representation and bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, 2008, 24, p.309-318.

[114] Winchester L, Yau C, Ragoussis J. Comparing cnv detection methods for snp arrays. *Briefings in Functional Genomics & Proteomics*, 2009, 8, p.353-366.

[115] Heun R, Kölsch H, Ibrahim-Verbaas CA, Combarros O, Aulchenko YS, Breteler M et al.. Interactions between ppar-α and inflammation-related cytokine genes on the development of alzheimer's disease, observed by the epistasis project. *International Journal of Molecular Epidemiology and Genetics*, 2012, 3, p.39-47.

[116] Kölsch H, Lehmann DJ, Ibrahim-Verbaas CA, Combarros O, van Duijn CM, Hammond N et al.. Interaction of insulin and ppar-α genes in alzheimer's disease: the epistasis project. *Journal of Neural Transmission*, 2012, 119, p.473-479.

[117] Brorsson C, Hansen NT, Lage K, Bergholdt R, Brunak S, Pociot F. Identification of t1d susceptibility genes within the mhc region by combining protein interaction networks and snp genotyping data. *Diabetes, Obesity & Metabolism*, 2009, 11 Suppl 1, p.60-66.

[118] Moore JH, Gilbert JC, Tsai C, Chiang F, Holden T, Barney N et al.. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 2006, 241, p.252-261.

[119] Zhang Y, Jiang B, Zhu J, Liu JS. Bayesian models for detecting epistatic interactions from genetic data. *Annals of Human Genetics*, 2011, 75, p.183-193.

[120] Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 2007, 39, p.1167-1173.

[121] Kooperberg C, Ruczinski I. Identifying interacting snps using monte carlo logic regression. *Genetic Epidemiology*, 2005, 28, p.157-170.

[122] Schwender H, Ickstadt K. Identification of snp interactions using logic regression. *Biostatistics*, 2008, 9, p.187-198.

[123] Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N et al.. Detection of gene x gene interactions in genome-wide association studies of human population data. *Human Heredity*, 2007, 63, p.67-84.

[124] Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD. A comparison of analytical methods for genetic association studies. *Genetic Epidemiology*, 2008, 32, p.767-778.

[125] Chen L, Yu G, Langefeld CD, Miller DJ, Guy RT, Raghuram J et al.. Comparative analysis of methods for detecting interacting loci. *BMC Genomics*, 2011, 12, p.344.

[126] Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 2000, 28, p.27-30.

[127] Holmans P, Green EK, Pahwa JS, Ferreira MAR, Purcell SM, Sklar P et al.. Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics*, 2009, 85, p.13-24.

[128] Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, 2007, 81, p.1278-1283.

[129] Zhang K, Cui S, Chang S, Zhang L, Wang J. I-gsea4gwas: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Research*, 2010, 38, p.W90-5.

[130] Abraham C, Cho J. Interleukin-23/th17 pathways and inflammatory bowel disease. *Inflammatory Bowel Diseases*, 2009, 15, p.1090-1100.

[131] Dong C. Th17 cells in development: an updated view of their molecular identity and genetic programming. *Nature Reviews. Immunology*, 2008, 8, p.337-348.

[132] Travis RC, Reeves GK, Green J, Bull D, Tipper SJ, Baker K et al.. Gene-environment interactions in 7610 women with breast cancer: prospective evidence from the million women study. *Lancet*, 2010, 375, p.2143-2151.

[133] Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H et al.. Influence of life stress on depression: moderation by a polymorphism in the 5-htt gene. *Science*, 2003, 301, p.386-389.

[134] Ku CS, Loy EY, Pawitan Y, Chia KS. The pursuit of genome-wide association studies: where are we now?. *Journal of Human Genetics*, 2010, 55, p.195-206.

[135] Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nature Reviews. Genetics*, 2011, 12, p.529-541.

[136] Sebastiani P, Timofeev N, Dworkis DA, Perls TT, Steinberg MH. Genome-wide association studies and the genetic dissection of complex traits. *American Journal of Hematology*, 2009, 84, p.504-515.

[137] Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 2011, 187, p.367-383.

[138] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of gwas discovery. *American Journal of Human Genetics*, 2012, 90, p.7-24.

[139] Manolio TA, Collins FS. The hapmap and genome-wide association studies in diagnosis and therapy. *Annual Review of Medicine*, 2009, 60, p.443-456.

[140] McClellan J, King M. Genetic heterogeneity in human disease. *Cell*, 2010, 141, p.210-217.

[141] McClellan JM, Susser E, King M. Schizophrenia: a common disease caused by multiple rare alleles. *The British Journal of Psychiatry : the Journal of Mental Science*, 2007, 190, p.194-199.

[142] Craddock N, O'Donovan MC, Owen MJ. Phenotypic and genetic complexity of psychosis. invited commentary on ... schizophrenia: a common disease caused by multiple rare alleles. *The British Journal of Psychiatry : the Journal of Mental Science*, 2007, 190, p.200-203.

[143] Gibson G. Rare and common variants: twenty arguments. *Nature Reviews. Genetics*, 2011, 13, p.135-145.

[144] Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT et al.. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 2010, 328, p.636-639.

[145] Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Human Molecular Genetics*, 2010, 19, p.4313-4318.

[146] Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DCY, Nazareth L et al.. Whole-genome sequencing in a patient with charcot-marie-tooth neuropathy. *The New England Journal of Medicine*, 2010, 362, p.1181-1191.

[147] Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G et al.. Performance comparison of exome dna sequencing technologies. *Nature Biotechnology*, 2011, 29, p.908-914.

[148] Singleton AB. Exome sequencing: a transformative technology. *Lancet Neurology*, 2011, 10, p.942-946.

[149] Pagnamenta AT, Lise S, Harrison V, Stewart H, Jayawant S, Quaghebeur G et al.. Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *Journal of Human Genetics*, 2012, 57, p.70-72.

[150] Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L et al.. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genetics*, 2011, 43, p.860-863.

[151] Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S et al.. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature Genetics*, 2011, 43, p.864-868.

[152] Chahrour MH, Yu TW, Lim ET, Ataman B, Coulter ME, Hill RS et al.. Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genetics*, 2012, 8, p.e1002635.

[153] O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP et al.. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 2012, 485, p.246-250.

[154] O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S et al.. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 2012, 44, p.471.

[155] Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ et al.. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 2012, en presse.

[156] Snape K, Ruark E, Tarpey P, Renwick A, Turnbull C, Seal S et al.. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Research and Treatment,* 2012, en presse.

[157] McGuire MM, Yatsenko A, Hoffner L, Jones M, Surti U, Rajkovic A. Whole exome sequencing in a random sample of north american women with leiomyomas identifies med12 mutations in majority of uterine leiomyomas. *PLoS One*, 2012, 7, p.e33251.

[158] Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence-based psychiatric genetics, aka the false dichotomy between common and rare variant hypotheses. *Molecular Psychiatry*, 2012, 17, p.474-485.

[159] Slatkin M. Exchangeable models of complex inherited diseases. *Genetics*, 2008, 179, p.2253-2261.

Liste des publications

1. **Taing L,** Le Clerc S, Ezzedine K, Latreille J, Labib T, Coulonges C, Bernard A, Melak S, Carpentier W, Malvy D, Jdid R, Galan P, Serge Hercberg S, Guinot C, Tschachler E, Zagury J. A genome-wide association study indentifies a gene associated with facial photoageing in caucasian woman. *Journal of Investigative Dermatology*, soumis.

2. **Taing L**, Coulonges C, Le Clerc S, Limou S, Dina C, Froguel P, Montes M, Spadoni J, Cantalloube H, Zagury J, Delaneau O. Computation by entropy of the information contained in snp datasets and applications. *BMC Genomics*, soumis.

3. Limou S, Delaneau O, van Manen D, An P, Sezgin E, Le Clerc S, Coulonges C, Troyer JL, Veldink JH, van den Berg LH, Spadoni J, **Taing L**, Labib T, Montes M, Delfraissy J, Schachter F, O'Brien SJ, Buchbinder S, van Natta ML, Jabs DA, Froguel P, Schuitemaker H, Winkler CA, Zagury J. Multicohort genomewide association study reveals a new signal of protection against hiv-1 acquisition. *The Journal of Infectious Diseases*, 2012, 205, p.1155-1162.

4. Le Clerc S, Coulonges C, Delaneau O, Van Manen D, Herbeck JT, Limou S, An P, Martinson JJ, Spadoni J, Therwath A, Veldink JH, van den Berg LH, **Taing L**, Labib T, Mellak S, Montes M, Delfraissy J, Schächter F, Winkler C, Froguel P, Mullins JI, Schuitemaker H, Zagury J. Screening low-frequency snps from genome-wide association study reveals a new risk allele for progression to aids. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 2011, 56, p.279-284.

5. Limou S, Coulonges C, Herbeck JT, van Manen D, An P, Le Clerc S, Delaneau O, Diop G, **Taing L**, Montes M, van't Wout AB, Gottlieb GS, Therwath A, Rouzioux C, Delfraissy J, Lelièvre J, Lévy Y, Hercberg S, Dina C, Phair J, Donfield S, Goedert JJ, Buchbinder S, Estaquier J, Schächter F, Gut I, Froguel P, Mullins JI, Schuitemaker H, Winkler C, Zagury J. Multiple-cohort genetic association study reveals cxcr6 as a new chemokine receptor involved in long-term nonprogression to aids. *The Journal of Infectious Diseases*, 2010, 202, p.908-915.

6. Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, Delaneau O, Labib T, **Taing L**, Sladek R, Deveau C, Ratsimandresy R, Montes M, Spadoni J, Lelièvre J, Lévy Y, Therwath A, Schächter F, Matsuda F, Gut I, Froguel P, Delfraissy J, Hercberg S, Zagury J. Genomewide association study of an aids-nonprogression cohort emphasizes the role played by hla genes

(anrs genomewide association study 02). *The Journal of Infectious Diseases*, 2009, 199, p.419-426.

7. Le Clerc S, Limou S, Coulonges C, Carpentier W, Dina C, **Taing L**, Delaneau O, Labib T, Sladek R, , Deveau C, Guillemain H, Ratsimandresy R, Montes M, Spadoni J, Therwath A, Schächter F, Matsuda F, Gut I, Lelièvre J, Lévy Y, Froguel P, Delfraissy J, Hercberg S, Zagury J. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for aids (anrs genomewide association study 03). *The Journal of Infectious Diseases*, 2009, 200, p.1194-1201.

Liste des communications orales

1. Lieng Taing. Entropy and Genomics.
4th Ermenonville International Workshop. 2009

2. Lieng Taing. Entropy: A simple way to tackle the bonferroni threshold.
5th Ermenonville International Workshop. 2010

3. Lieng Taing. Preliminary results from the Photoageing GWAS project.
6th Ermenonville International Workshop. 2011

Lieng TAING

Approches bioinformatiques pour l'exploitation des données génomiques

Résumé

Les technologies actuelles permettent d'explorer le génome entier pour identifier des variants génétiques associés à des phénotypes particuliers, notamment de maladies. C'est le rôle de la bioinformatique de répondre à cette problématique.

Dans le cadre de cette thèse, un nouvel outil logiciel a été développé afin de mesurer avec une précision le nombre de marqueurs génétiques effectivement indépendants correspondant à un ensemble de marqueurs géntotypés dans une population donnée en utilisant l'entropie de Shannon. Ce calcul peut avoir de nombreuses applications pour l'exploitation des données génomiques.

Ce travail de recherche porte aussi sur une étude génome-entier sur le photo-vieillissement. Dans cette étude, réalisée sur 502 femmes génotypées deux gènes (*STXBP5L* et *FBX040*) associés à un SNP passant le seuil de Bonferroni ont été identifiés, dont l'implication dans le photo-vieillissement était jusqu'alors inconnue.

Mots clés : études d'association, entropie de Shannon, photo-vieillissement, SNP, tests multiples

Résumé en anglais

New technologies allow the exploration of the whole genome to identify genetic variants associated with verious phenotypes, in particular diseases. Bioinformatics aims at helping to answer these questions.

In the context of my PhD thesis, I have first developed a new software allowing to measure with a good precision the number of really independent genetic markers present in a set of markers genotyped in a given population using Shannon's entropy. This computation may have several applications for the exploitation of genomic data.

I have also completed a genome-wide association study on photo-ageing. In this study of 502 women genotyped with OmniOne Illumina chips (1M SNPs), I have identified two genes (*STXBP5L* et *FBX040*) associated with a SNP that passes the Bonferroni threshold, whose implication in photo-ageing was not suspected until now.

Keywords: GWAS, multitesting, photoageing, Shannon's entropy, SNP